

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**



**PCT**

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup>:</b> <b>C12Q 1/68, C07H 21/04</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 97/13877</b> <b>(43) International Publication Date:</b> 17 April 1997 (17.04.97)
<b>(21) International Application Number:</b> PCT/US96/16342 <b>(22) International Filing Date:</b> 11 October 1996 (11.10.96) <b>(30) Priority Data:</b> PCT/US95/12791 12 October 1995 (12.10.95) WO (34) Countries for which the regional or international application was filed: US et al. PCT/US96/09513 6 June 1996 (06.06.96) WO (34) Countries for which the regional or international application was filed: US et al.  <b>(60) Parent Application or Grant</b> (63) Related by Continuation US Not furnished (CIP) Filed on Not furnished  <b>(71) Applicant (for all designated States except US):</b> LYNX THERAPEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US).  <b>(72) Inventor; and</b> <b>(75) Inventor/Applicant (for US only):</b> MARTIN, David, W. [US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).		<b>(74) Agent:</b> POWERS, Vincent, M.; Dehlinger & Associates, Post Office Box 60850, Palo Alto, CA 94306-0850 (US).  <b>(81) Designated States:</b> AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV, NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  <b>Published</b> <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
<b>(54) Title:</b> MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION  <b>(57) Abstract</b>  A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

5 Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates.

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year, and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is estimated that the pharmaceutical industry spends an average of about 300 million dollars to bring a new pharmaceutical compound to market, e.g. Biotechnology, 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is, as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example, through relatively inexpensive binding assays or in vitro screening assays, to pharmacokinetic studies, to toxicity studies, to efficacy studies in model systems, to preliminary clinical studies, and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

- 1 -

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical is administered for a substantial portion of the lifetime of the test animal. In the case of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993); Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols, 5 5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

10 The availability of new assays for assessing the toxicity of compounds, such as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

15

#### Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems.

20

Another object of the invention is to provide a database on which to base decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

25

A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

30

Another object of the invention is to identify genes whose expression is predictive of deleterious toxicity.

35

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

#### Brief Description of the Drawings

Figure 1 is a flow chart representation of an algorithm for generating minimally cross-hybridizing sets of oligonucleotides.

Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

#### Definitions

"Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides, anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

5 "Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means  
10 that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse  
15 Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analog" in reference to  
20 nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990). or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

25 As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides,  
30 usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide, e.g. cytosines, within the target polynucleotide  
35 "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ..." for "C-(not C)-(not C)-C-(not C)-C ..." and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

#### Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the above method.

#### Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Essentials of Toxicology, 4th Ed. (Academic Press, New York, 1996); Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier, editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

5 In toxicity testing, two groups of test organisms are usually employed: one group serves as a control and the other group receives the test compound in a single dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity tests). Since in most cases, the extraction of tissue as called for in the method of the invention requires sacrificing the test animal, both the control group and the group receiving compound must be large enough to permit removal of animals for sampling  
10 tissues, if it is desired to observe the dynamics of gene expression through the duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature for selecting the appropriate test organism for the compound being tested, route of administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl  
15 in water) is the solute of choice for the test compound since these solvents permit administration by a variety of routes. When this is not possible because of solubility limitations, it is necessary to resort to the use of vegetable oils such as corn oil or even organic solvents, of which propylene glycol is commonly used. Whenever possible the use of suspension or emulsion should be avoided except for oral  
20 administration. Regardless of the route of administration, the volume required to administer a given dose is limited by the size of the animal that is used. It is desirable to keep the volume of each dose uniform within and between groups of animals. When rats or mice are used the volume administered by the oral route should not exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline  
25 solutions are used for parenteral injection the volumes that are tolerated are limited, although such solutions are ordinarily thought of as being innocuous. The intravenous LD<sub>50</sub> of distilled water in the mouse is approximately 0.044 ml per gram and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for  
30 generating test atmospheres are necessary. Dose estimation becomes very complicated. The methods usually involve aerosolization or nebulization of fluids containing the compound. If the agent to be tested is a fluid that has an appreciable vapor pressure, it may be administered by passing air through the solution under controlled temperature conditions. Under these conditions, dose is estimated from the  
35 volume of air inhaled per unit time, the temperature of the solution, and the vapor pressure of the agent involved. Gases are metered from reservoirs. When particles of a solution are to be administered, unless the particle size is less than about 2  $\mu$ m the particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

5            Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of  
10 administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

15            Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomatology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses of the compound in a single species. This necessitates selection of a route of  
20 administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test  
25 organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without  
30 inducing significant harm to the test animals. Preferably, in the method of the invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

35

Hematology	Blood Chemistry	Urine Analyses
erythrocyte count	sodium	pH
total leukocyte count	potassium	specific gravity
differential leukocyte count	chloride	total protein
hematocrit	calcium	sediment
hemoglobin	carbon dioxide	glucose
	serum glutamine-pyruvate transaminase	ketones
	serum glutamin-oxalacetic transaminase	bilirubin
	serum protein	
	electrophoresis	
	blood sugar	
	blood urea nitrogen	
	total serum protein	
	serum albumin	
	total serum bilirubin	

5

#### Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting, tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For example, a minimally cross-hybridizing set may consist of a set of individually synthesized 10-mer sequences that differ from each other by at least 4 nucleotides. such set having a maximum size of 332 (when composed of 3 kinds of nucleotides and counted using a computer program such as disclosed in Appendix Ic). Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of  $9^3$ , or 729, 12-mers. The number 9 is number of oligonucleotides listed by the computer program of Appendix Ia, which assumes, as with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set is described as "maximal" because the computer programs of Appendices Ia-c provide the largest set for a given input (e.g. length, composition, difference in number of nucleotides between members). Additional minimally cross-hybridizing sets may be formed from subsets of such calculated sets.

Oligonucleotide tags may be single stranded and be designed for specific hybridization to single stranded tag complements by duplex formation or for specific hybridization to double stranded tag complements by triplex formation. Oligonucleotide tags may also be double stranded and be designed for specific hybridization to single stranded tag complements by triplex formation.

When synthesized combinatorially, an oligonucleotide tag preferably consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. The number is generally much less than the number of all possible sequences the length of the tag, which for a tag  $n$  nucleotides long would be  $4^n$ .

Complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences. As used herein in reference to oligonucleotide tags and tag complements, the term "repertoire" means the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed  
 5 under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those  
 10 exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of  
 15 minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

20

Table I

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set	Maximal Size of Minimally Cross-Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	3	9	6561	$5.90 \times 10^4$
6	3	27	$5.3 \times 10^5$	$1.43 \times 10^7$
7	4	27	$5.3 \times 10^5$	$1.43 \times 10^7$
7	5	8	4096	$3.28 \times 10^4$
8	3	190	$1.30 \times 10^9$	$2.48 \times 10^{11}$
8	4	62	$1.48 \times 10^7$	$9.16 \times 10^8$
8	5	18	$1.05 \times 10^5$	$1.89 \times 10^6$
9	5	39	$2.31 \times 10^6$	$9.02 \times 10^7$
10	5	332	$1.21 \times 10^{10}$	
10	6	28	$6.15 \times 10^5$	$1.72 \times 10^7$
11	5	187		
18	6	$\approx 25000$		

18

12

24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and selected from the same minimally cross-hybridizing set. For oligonucleotides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table  $M_n$ ,  $n=1$ , is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit  $S_1$  is selected and compared (120) with successive subunits  $S_i$  for  $i=n+1$  to the end of the table. Whenever a successive subunit has the required number of mismatches to be a member of the minimally cross-hybridizing set, it is saved in a new table  $M_{n+1}$  (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons,  $M_2$  will contain  $S_1$ ; in the second set of comparisons,  $M_3$  will contain  $S_1$  and  $S_2$ ; in the third set of comparisons,  $M_4$  will contain  $S_1$ ,  $S_2$ , and  $S_3$ ; and so on. Similarly, comparisons in table  $M_j$  will be between  $S_j$  and all successive subunits in  $M_j$ . Note that each successive table  $M_{n+1}$  is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table  $M_n$  has been compared (140) the old table is replaced by the new table  $M_{n+1}$ , and the next round of comparisons are begun. The process stops (160) when a table  $M_n$  is reached that contains no successive subunits to compare to the selected subunit  $S_i$ , i.e.  $M_n=M_{n+1}$ .

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991); and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, there are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the effects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in *italic* below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar terminal "word" on any other tag complement. Such an augmented tag would have the form:

<i>W</i>	<i>W</i> <sub>1</sub>	<i>W</i> <sub>2</sub>	...	<i>W</i> <sub><i>k</i>-1</sub>	<i>W</i> <sub><i>k</i></sub>	<i>W</i>
<i>W</i> '	<i>W</i> <sub>1</sub> '	<i>W</i> <sub>2</sub> '	...	<i>W</i> <sub><i>k</i>-1</sub> '	<i>W</i> <sub><i>k</i></sub> '	<i>W</i> '

where the primed *W*'s indicate complements. With ends of tags always forming perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5'→3' exonuclease activity of a DNA polymerase. The following is an exemplary minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

5

Table II

Word:	w <sub>1</sub>	w <sub>2</sub>	w <sub>3</sub>	w <sub>4</sub>
Sequence:	GATT	TGAT	TAGA	TTTG
Word:	w <sub>5</sub>	w <sub>6</sub>	w <sub>7</sub>	w <sub>8</sub>
Sequence:	GTAA	AGTA	ATGT	AAAG

10 In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table III. Clearly, additional sets can be generated by substituting different groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

15

Table III

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

<u>Set 1</u>	<u>Set 2</u>	<u>Set 3</u>	<u>Set 4</u>	<u>Set 5</u>	<u>Set 6</u>
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

<u>Set 7</u>	<u>Set 8</u>	<u>Set 9</u>	<u>Set 10</u>	<u>Set 11</u>	<u>Set 12</u>
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al. International patent application PCT/US93/03418 or Lyttle et al, Biotechniques, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags.

- 5 That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle  
10 of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

- Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a  
15 single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags  
20 may then be inserted into cloning vectors along with target polynucleotides for sorting and manipulation of the target polynucleotide in accordance with the invention.

- When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide N3'→P5' phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate  
25 complements, as an alternative to the "stripping" reaction employing the 3'→5' exonuclease activity of a DNA polymerase to render a tag single stranded.

- Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length  
30 from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV  
Numbers of Subunits in Tags in Preferred Embodiments

35

<u>Monomers in Subunit</u>	<u>Nucleotides in Oligonucleotide Tag</u>		
	(12-60)	(18-40)	(25-40)

3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

- 5 Preferably, repertoires of single stranded oligonucleotide tags of the invention contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

### Triplex Tags

- 10 In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A\*T or C-G\*C motifs (where "-" indicates Watson-Crick pairing and "\*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl. Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan, Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and the like. Conditions for annealing single-stranded or duplex tags to their single-stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like. Use of triplex tags has the advantage of not requiring a "stripping" reaction with polymerase to expose the tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex hybridization are double stranded DNA and the corresponding tag complements are single stranded. More preferably, 5-methylcytosine is used in place of cytosine in the tag complements in order to broaden the range of pH stability of the triplex formed between a tag and its complement. Preferred conditions for forming triplexes are fully disclosed in the above references. Briefly, hybridization takes place in concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at pH below 5.5 ( or 6.5 if 5-methylcytosine is employed). Hybridization temperature depends on the length and composition of the tag; however, for an 18-20-mer tag of longer, hybridization at room temperature is adequate. Washes may be conducted with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM MgCl<sub>2</sub>, pH 5.8, at room temperature. Tags may be eluted from their tag complements by incubation in a similar salt solution at pH 9.0.

Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes may be generated by the computer program of Appendix Ic, or similar programs. An exemplary set of double stranded 8-mer words are listed below in capital letters with the corresponding complements in small letters. Each such word differs from each of the other words in the set by three base pairs.

25

Table V  
Exemplary Minimally Cross-Hybridizing  
Set of DoubleStranded 8-mer Tags

5' -AAGGAGAG	5' -AAAGGGGA	5' -AGAGAAGA	5' -AGGGGGGG
3' -TTCCTCTC	3' -TTTCCCT	3' -TCTCTTCT	3' -TCCCCCCC
3' -ttcctctc	3' -tttcccct	3' -tctcttct	3' -tccccccc
5' -AAAAAAA	5' -AAGAGAGA	5' -AGGAAAAG	5' -GAAAGGAG
3' -TTTTTTT	3' -TTCTCTCT	3' -TCCTTTTC	3' -CTTCTCTC
3' -tttttttt	3' -ttctctct	3' -tccttttc	3' -cttctctc
5' -AAAAAGGG	5' -AGAAGAGG	5' -AGGAAGGA	5' -GAAGAAGG
3' -TTTTTCCC	3' -TCTTCTCC	3' -TCCTTCCT	3' -CTTCTTCC
3' -tttttccc	3' -tcttctcc	3' -tccttcct	3' -cttcttcc
5' -AAAGGAAG	5' -AGAAGGAA	5' -AGGGGAAA	5' -GAAGAGAA
3' -TTTCCTTC	3' -TCTTCCTT	3' -TCCCCTTT	3' -CTTCTCTT
3' -tttccttc	3' -tcttcctt	3' -tccccttt	3' -cttctctt

5

10

Table VI  
Repertoire Size of Various Double Stranded Tags  
That Form Triplexes with Their Tag Complements

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	2	8	4096	$3.2 \times 10^4$
6	3	8	4096	$3.2 \times 10^4$
8	3	16	$6.5 \times 10^4$	$1.05 \times 10^6$
10	5	8	4096	
15	5	92		
20	6	765		
20	8	92		
20	10	22		

15 Preferably, repertoires of double stranded oligonucleotide tags of the invention contain at least 10 members; more preferably, repertoires of such tags contain at least 100 members. Preferably, words are between 4 and 8 nucleotides in length for combinatorially synthesized double stranded oligonucleotide tags, and oligonucleotide tags are between 12 and 60 base pairs in length. More preferably, such tags are  
20 between 18 and 40 base pairs in length.

#### Solid Phase Supports

25 Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several  $\mu\text{m}^2$ , e.g. 3-5, to several hundred  $\mu\text{m}^2$ , e.g. 100-500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, *Nucleic Acids Research*, 16: 10861-10880 (1988); Albretsen et al, *Anal. Biochem.*, 189: 40-50 (1990); Wolf et al, *Nucleic Acids Research*, 15: 2911-2926 (1987); or Ghosh et al, *Nucleic Acids Research*, 15: 5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: *Meth. Enzymol.*, Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046,720; and Pon, Chapter 19, in Agrawal, editor, *Methods in Molecular Biology*, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel<sup>TM</sup>, Rapp Polymere, Tübingen Germany); and the like. Selection of the support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumental advantages  
5 when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, *Biotechniques*, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, *J. Chem. Soc. Commun.*, 1989: 891-893; Damha et  
10 al, *Nucleic Acids Research*, 18: 3813-3821 (1990); Beattie et al, *Clinical Chemistry*, 39: 719-722 (1993); Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag  
15 complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, *Proc. Natl. Acad. Sci.*, 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, *Genomics*, 13:  
20 1008-1017 (1992); and Maskos and Southern, *Nucleic Acids Research*, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces  
25 are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000  $\mu\text{m}$  diameter are preferable, as they facilitate the  
30 construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached.  
35 e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide. Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5  $\mu$ m diameter GMA beads are employed.

10

Attaching Tags to Polynucleotides  
For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably, the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing primers has a complexity of  $8^9$ , or about  $1.34 \times 10^8$ . As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about  $5 \times 10^5$  copies of mRNA molecules of about  $3.4 \times 10^4$  different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about  $5 \times 10^7$  mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1% efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about  $10^8$  cDNA clones from mRNA extracted from  $10^6$  mammalian cells.

Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

If mRNA were extracted from  $10^6$  cells (which would correspond to about 0.5  $\mu\text{g}$  of poly(A)<sup>+</sup> RNA), and if primers were present in about 10-100 fold concentration excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10  $\mu\text{L}$  1.8 kb mRNA at 1 mg/mL equals about  $1.68 \times 10^{-11}$  moles and 10  $\mu\text{L}$  18-mer primer at 1 mg/mL equals about  $1.68 \times 10^{-9}$  moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about  $5 \times 10^{11}$  vectors containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector--occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

If a sample of  $n$  tag-polynucleotide conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution,  $P(r) = e^{-\lambda} (\lambda)^r / r!$ , where  $r$  is the number of conjugates having the same tag and  $\lambda = np$ , where  $p$  is the probability of a given tag being selected. If  $n = 10^6$  and  $p = 1/(1.34 \times$

$10^8$ ), then  $\lambda = .00746$  and  $P(2) = 2.76 \times 10^{-5}$ . Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the  $5 \times 10^{11}$  mRNAs are perfectly converted into  $5 \times 10^{11}$  vectors with tag-cDNA conjugates as inserts and that the  $5 \times 10^{11}$  vectors are in a reaction solution having a volume of 100  $\mu$ l. Four 10-fold serial dilutions may be carried out by transferring 10  $\mu$ l from the original solution into a vessel containing 90  $\mu$ l of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100  $\mu$ l solution containing  $5 \times 10^5$  vector molecules per  $\mu$ l. A 2  $\mu$ l aliquot from this solution yields  $10^6$  vectors containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient. Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of  $10^6$  conjugates.

A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool, and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the polynucleotides have unique tags attached. More preferably, it means that at least ninety percent of the polynucleotides have unique tags attached. Still  
 5 more preferably, it means that at least ninety-five percent of the polynucleotides have unique tags attached. And, most preferably, it means that at least ninety-nine percent of the polynucleotides have unique tags attached.

Preferably, when the population of polynucleotides consists of messenger RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the  
 10 mRNA with a set of primers preferably containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

5' -mRNA- [A]<sub>n</sub> -3'  
 15 [T]<sub>19</sub>GG[W,W,W,C]<sub>9</sub>ACCAGCTGATC-5' -biotin

where "[W,W,W,C]<sub>9</sub>" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences  
 20 listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

25 5' -[G,W,W,W]<sub>9</sub>TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion, and the second strand of the cDNA is synthesized using, for example, a primer of the following form (SEQ ID NO: 2):

30 5' -NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst YI restriction  
 35 site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst YI and Sal I digestion, the exemplary conjugate would have the form:

5'-RCGACCA [C, W, W, W] 9GG [T] 19- cDNA -NNNR  
GGT [G, W, W, W] 9CC [A] 19- rDNA -NNNYCTAG-5'

The polynucleotide-tag conjugates may then be manipulated using standard molecular biology techniques. For example, the above conjugate--which is actually a mixture--may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning System (La Jolla, CA); transfected into a host, such as a commercially available host bacteria; which is then cultured to increase the number of conjugates. The cloning vectors may then be isolated using standard techniques, e.g. Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Alternatively, appropriate adaptors and primers may be employed so that the conjugate population can be increased by PCR.

Preferably, when the ligase-based method of sequencing is employed, the Bst YI and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites (SEQ ID NO: 3):

5' -GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'  
FokI BamHI XhoI

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

Tags can be conjugated to cDNAs of existing libraries by standard cloning methods. cDNAs are excised from their existing vector, isolated, and then ligated into a vector containing a repertoire of tags. Preferably, the tag-containing vector is linearized by cleaving with two restriction enzymes so that the excised cDNAs can be ligated in a predetermined orientation. The concentration of the linearized tag-containing vector is in substantial excess over that of the cDNA inserts so that ligation provides an inherent sampling of tags.

A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'→3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single deoxynucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double stranded fragment until a complement of the single deoxynucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the 5'→3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, *Critical Reviews in Biochemistry and Molecular Biology*, 26: 227-259 (1991); Sambrook et al, *Molecular Cloning: A Laboratory Manual*, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50  $\mu\text{m}$  are loaded with about  $10^5$  polynucleotides, and GMA beads of diameter in the range of 5-10  $\mu\text{m}$  are loaded with a few tens of thousand of polynucleotides, e.g.  $4 \times 10^4$  to  $6 \times 10^4$ .

- In the preferred embodiment, tag complements are synthesized on
- 5 microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance
- 10 for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

15

Microparticle diameter	5 $\mu\text{m}$	10 $\mu\text{m}$	20 $\mu\text{m}$	40 $\mu\text{m}$
Max. no. polynucleotides loaded at 1 per $10^5$ sq. angstrom		$3 \times 10^5$	$1.26 \times 10^6$	$5 \times 10^6$
Approx. area of monolayer of $10^6$ microparticles	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

- 20 The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

25

Table VII

Number of microparticles in sample (as fraction of repertoire size), $m$	Fraction of repertoire of tag complements present in sample, $1-e^{-m}$	Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$	Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

#### High Specificity Sorting and Panning

5        The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

10        Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of  $8^8$ , or about  $1.67 \times 10^7$ , tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of  $1.67 \times 10^7$  microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to, the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by reduction, e.g. Herman et al, *Anal. Biochem.*, 156: 48-55 (1986), or that is cleaved photochemically, e.g. Olejnik et al, *Nucleic Acids Research*, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

25 5'-RCGACCA[C,W,W,W]<sub>9</sub>GG[T]<sub>19</sub>- cDNA -NNNR  
GGT[G,W,W,W]<sub>9</sub>CC[A]<sub>19</sub>- rDNA -NNNYCTAG-5'

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:

30  
5'-XXXXXXXXXXXXXXXXXXXX  
XXXXXXXXXXXXXXXXXXXXYGAT

35 Right Adapter

40 GATCZZ**ACTAGT**ZZZZZZZZZZZZ-3'  
ZZ**TGATCA**ZZZZZZZZZZZZ

## Left Adapter

ZZTGATCAZZZZZZZZZZZZ-5'-biotin

5

## Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage ready for single base sequencing), and the X's and Z's are nucleotides selected so that the annealing and dissociation temperatures of the respective primers are approximately the same. After ligation of the adapters and amplification by PCR using the biotinylated primer, the tags of the conjugates are rendered single stranded by the exonuclease activity of T4 DNA polymerase and conjugates are combined with a sample of microparticles, e.g. a repertoire equivalent, with tag complements attached. After annealing under stringent conditions (to minimize mis-attachment of tags), the conjugates are preferably ligated to their tag complements and the loaded microparticles are separated from the unloaded microparticles by capture with avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about 10,500 ( $=16,700 \times .63$ ) loaded microparticles with different tags, which may be released from the magnetic beads by cleavage with Spe I. By repeating this process 40-50 times with new samples of microparticles and tag-cDNA conjugates,  $4-5 \times 10^5$  cDNAs can be accumulated by pooling the released microparticles. The pooled microparticles may then be simultaneously sequenced by a single-base sequencing technique.

Determining how many times to repeat the sampling and panning steps--or more generally, determining how many cDNAs to analyze, depends on one's objective. If the objective is to monitor the changes in abundance of relatively common sequences, e.g. making up 5% or more of a population, then relatively small samples, i.e. a small fraction of the total population size, may allow statistically significant estimates of relative abundances. On the other hand, if one seeks to monitor the abundances of rare sequences, e.g. making up 0.1% or less of a population, then large samples are required. Generally, there is a direct relationship between sample size and the reliability of the estimates of relative abundances based on the sample. There is extensive guidance in the literature on determining appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al, *Nucleic Acids Research*, 23:185-191 (1994); Good, *Biometrika*, 40: 16-264 (1953); Bunge et al, *J. Am. Stat. Assoc.*, 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing  $10^5$  to  $10^8$  independent clones of  $3.0$ - $3.5 \times 10^4$  different sequences, a sample of at least  $10^4$  sequences are accumulated for analysis of each library. More preferably, a sample of at least  $10^5$  sequences are accumulated for the analysis of each library; and most preferably, a sample of at least  $5 \times 10^5$  sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

10

#### Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a time, in a sequence through successive cycles of treatment and detection. Such methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of probes that are applied sequentially to the plurality of target polynucleotides on the different microparticles. An exemplary set of such probes are shown below:

5

Set 1	Set 2	Set 3	Set 4
ANNNN...NN N...NNTT...T*	dANNNN...NN d N...NNTT...T	dANNNN...NN N...NNTT...T	dANNNN...NN N...NNTT...T
dCNNNN...NN N...NNTT...T	CNNNN...NN N...NNTT...T*	dCNNNN...NN N...NNTT...T	dCNNNN...NN N...NNTT...T
dGNNNN...NN N...NNTT...T	dGNNNN...NN N...NNTT...T	GNNNN...NN N...NNTT...T*	dGNNNN...NN N...NNTT...T
dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	dTNNNN...NN N...NNTT...T	TNNNN...NN N...NNTT...T*

where each of the listed probes represents a mixture of  $4^3=64$  oligonucleotides such that the identity of the 3' terminal nucleotide of the top strand is fixed and the other positions in the protruding strand are filled by every 3-mer permutation of nucleotides, or complexity reducing analogs. The listed probes are also shown with a single stranded poly-T tail with a signal generating moiety attached to the terminal thymidine, shown as "T\*". The "d" on the unlabeled probes designates a ligation-blocking moiety or absence of 3'-hydroxyl, which prevents unlabeled probes from being ligated. Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment, the probes of set 1 are first applied to the plurality of target polynucleotides and treated with a ligase so that target polynucleotides having a thymidine complementary to the 3' terminal adenosine of the labeled probes are ligated. The unlabeled probes are simultaneously applied to minimize inappropriate ligations. The locations of the target polynucleotides that form ligated complexes with probes terminating in "A" are identified by the signal generated by the label carried on the probe. After washing and cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming ligated complexes with probes terminating in "C" are identified by location. Similarly, the probes of sets 3 and 4 are applied and locations of positive signals identified. This process of sequentially applying the four sets of probes continues until the desired number of nucleotides are identified on the target polynucleotides. Clearly, one of ordinary skill could construct similar sets of probes that could have many variations, such as having protruding strands of different lengths, different moieties to block ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available from Apple Computer (Cupertino, CA). Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic, e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336, whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD array.

The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100  $\mu\text{m}$ . Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g. from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per  $\text{cm}^2$ .

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

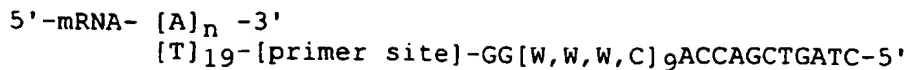
in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique  
5 restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the  
10 associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the  
15 initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type IIIs restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture  
- comprises about 10-15 percent of the biotinylated probe.

20 In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched  
25 in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

#### Identification of Novel Polynucleotides in cDNA Libraries

30 Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as  
35 taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may be hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):



where  $[W,W,W,C]_9$  represents a tag as described above, "ACCAGCTGATC" is an optional sequence forming a restriction site in double stranded form, and "primer site" is a sequence common to all members of the library that is later used as a primer binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional techniques, the double stranded fragments are inserted into a cloning vector as described above and amplified. The amplified library is then sampled and the sample amplified. The cloning vectors from the amplified sample are isolated, and the tagged cDNA fragments excised and purified. After rendering the tag single stranded with a polymerase as described above, the fragments are methylated and sorted onto microparticles in accordance with the invention. Preferably, as described above, the cloning vector is constructed so that the tagged cDNAs can be excised with an endonuclease, such as Fok I, that will allow immediate sequencing by the preferred single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library, or one or more large fractions of the library, in accordance with the invention until a sufficient number of nucleotides are identified on each cDNA for unique representation in the genome of the organism from which the library is derived. For example, if the library is derived from mammalian mRNA then a randomly selected sequence 14-15 nucleotides long is expected to have unique representation among the 2-3 thousand megabases of the typical mammalian genome. Of course identification of far fewer nucleotides would be sufficient for unique representation in a library derived from bacteria, or other lower organisms. Preferably, at least 20-30 nucleotides are identified to ensure unique representation and to permit construction of a suitable primer as described below. The tabulated sequences may then be compared to known sequences to identify unique cDNAs.

Unique cDNAs are then isolated by conventional techniques, e.g. constructing a probe from the PCR amplicon produced with primers directed to the prime site and the portion of the cDNA whose sequence was determined. The probe may then be used to identify the cDNA in a library using a conventional screening protocol.

The above method for identifying new cDNAs may also be used to fingerprint mRNA populations, either in isolated measurements or in the context of a dynamically changing population. Partial sequence information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of cDNAs attached to separate microparticles as described in the above method.

**Example 1****Construction of a Tag Library**

An exemplary tag library is constructed as follows to form the chemically  
 5 synthesized 9-word tags of nucleotides A, G, and T defined by the formula:



where "[ ${}^4\text{(A,G,T)}_9$ ]" indicates a tag mixture where each tag consists of nine 4-mer  
 10 words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the  
 following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):

5' - AGTGGCTGGGCATCGGACCG  
 TCACCGACCCGTAGCCp

5' - GGGGCCCAGTCAGCGTCGAT  
 GGGTCAGTCGCAGCTA

15

LEFT

RIGHT

The right and left primer binding regions are ligated to the above tag mixture, after  
 which the single stranded portion of the ligated structure is filled with DNA  
 20 polymerase then mixed with the right and left primers indicated below and amplified  
 to give a tag library (SEQ ID NO: 6).

**Left Primer**

25

5' - AGTGGCTGGGCATCGGACCG

5' - AGTGGCTGGGCATCGGACCG- [ ${}^4\text{(A,G,T)}_9$ ]-GGGGCCCAGTCAGCGTCGAT  
 TCACCGACCCGTAGCCTGGC- [ ${}^4\text{(A,G,T)}_9$ ]-CCCCGGGTCAGTCGCAGCTA

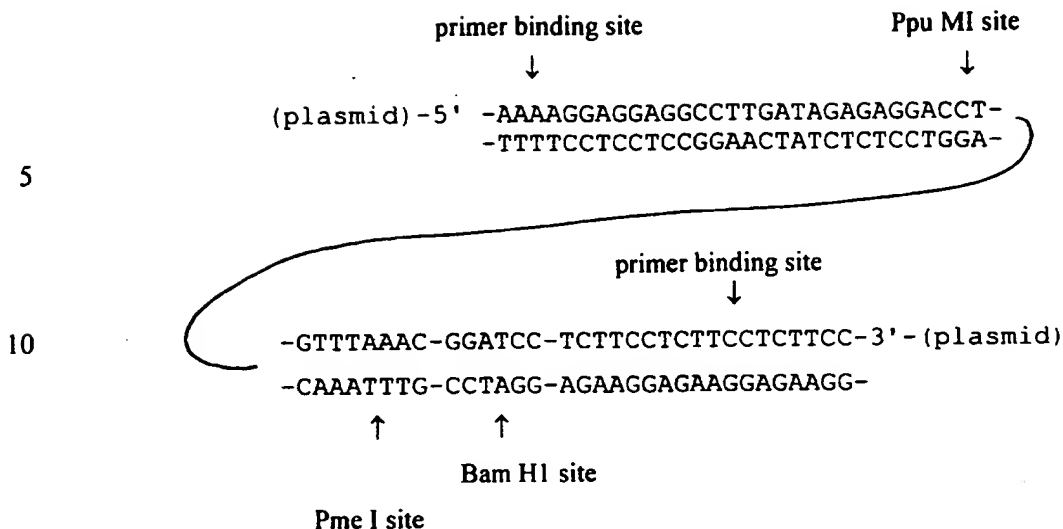
30

CCCCGGGTCAGTCGCAGCTA-5'

**Right Primer**

35 The underlined portion of the left primer binding region indicates a Rsr II recognition  
 site. The left-most underlined region of the right primer binding region indicates  
 recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I.  
 The right-most underlined region of the right primer binding region indicates the  
 recognition site for Hga I. Optionally, the right or left primers may be synthesized  
 40 with a biotin attached (using conventional reagents, e.g. available from Clontech  
 Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or  
 cleavage.

**NOT FURNISHED UPON FILING**



The plasmid is cleaved with Ppu MI and Pme I (to give a Rsr II-compatible end and a flush end so that the insert is oriented) and then methylated with DAM methylase. The tag-containing construct is cleaved with Rsr II and then ligated to the open plasmid, after which the conjugate is cleaved with Mbo I and Bam HI to permit ligation and closing of the plasmid. The plasmid is then amplified and isolated and used in accordance with the invention.

### Example 3

#### Changes in Gene Expression Profiles in Liver Tissue of Rats

##### Exposed to Various Xenobiotic Agents

In this experiment, to test the capability of the method of the invention to detect genes induced as a result of exposure to xenobiotic compounds, the gene expression profile of rat liver tissue is examined following administration of several compounds known to induce the expression of cytochrome P-450 isoenzymes. The results obtained from the method of the invention are compared to results obtained from reverse transcriptase PCR measurements and immunochemical measurements of the cytochrome P-450 isoenzymes. Protocols and materials for the latter assays are described in Morris et al, Biochemical Pharmacology, 52: 781-792 (1996).

Male Sprague-Dawley rats between the ages of 6 and 8 weeks and weighing 200-300 g are used, and food and water are available to the animals *ad lib*. Test compounds are phenobarbital (PB), metyrapone (MET), dexamethasone (DEX), clofibrate (CLO), corn oil (CO), and  $\beta$ -naphthoflavone (BNF), and are available from Sigma Chemical Co. (St. Louis, MO). Antibodies against specific P-450 enzymes are available from the following sources: rabbit anti-rat CYP3A1 from Human Biologics, Inc. (Phoenix, AZ); goat anti-rat CYP4A1 from Daiichi Pure Chemicals Co. (Tokyo,

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson

5 ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with  
10 H<sub>2</sub>O and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately  
15 100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 µg/ml. Total RNA is stored in  
- diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent  
20 degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 µg of poly(A)<sup>+</sup> RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto  
25 Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is 8<sup>6</sup> or about 2.6 x 10<sup>5</sup>. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about 10<sup>6</sup> 10 µm  
30 diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as  
35 described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about  $5 \times 10^4$  to  $2 \times 10^5$  tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency  
5 distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20  $\mu$ L reaction mixture is prepared containing 1x reverse  
10 transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5  $\mu$ M oligo d(T)<sub>15</sub> primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For  
15 PCR amplification of cDNA, a 10  $\mu$ L reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM MgCl<sub>2</sub>, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting,  
20 annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min., respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE  
25 gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50  $\mu$ g/lane and resolved under constant current (250 V) for approximately 4 hours at 2°C. Proteins are transferred to nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing  
30 120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed  
35 substantial agreement.

**APPENDIX Ia**  
**Exemplary computer program for generating**  
**minimally cross hybridizing sets**  
**(single stranded tag/single stranded tag complement)**

```
C
C
C
Program minxh
integer*2 subl(6),mset1(1000,6),mset2(1000,6)
dimension nbase(6)

C
C
write(*,*)'ENTER SUBUNIT LENGTH'
read(*,100)nsub
format(il)
open(1,file='sub4.dat',form='formatted',status='new')

C
C
nset=0
do 7000 m1=1,3
  do 7000 m2=1,3
    do 7000 m3=1,3
      do 7000 m4=1,3
        subl(1)=m1
        subl(2)=m2
        subl(3)=m3
        subl(4)=m4

C
C
ndiff=3

C
C
C
Generate set of subunits differing from
subl by at least ndiff nucleotides.
Save in mset1.

C
C
jj=1
do 900 j=1,nsub
  mset1(1,j)=subl(j)

C
C
do 1000 k1=1,3
  do 1000 k2=1,3
    do 1000 k3=1,3
      do 1000 k4=1,3

C
C
nbase(1)=k1
nbase(2)=k2
nbase(3)=k3
nbase(4)=k4
```

```

c
      n=0
      do 1200 j=1, nsub
        if (sub1(j).eq.1 .and. nbase(j).ne.1 .or.
1         sub1(j).eq.2 .and. nbase(j).ne.2 .or.
3         sub1(j).eq.3 .and. nbase(j).ne.3) then
          n=n+1
          endif
1200      continue
c
c
c      if (n.ge.ndiff) then
c
c
c          If number of mismatches
c          is greater than or equal
c          to ndiff then record
c          subunit in matrix mset
c
c
c          jj=jj+1
c          do 1100 i=1, nsub
1100      mset1(jj,i)=nbase(i)
c          endif
c
c
c      continue
1000
c
c      do 1325 j2=1, nsub
c      mset2(1,j2)=mset1(1,j2)
1325      mset2(2,j2)=mset1(2,j2)
c
c
c          Compare subunit 2 from
c          mset1 with each successive
c          subunit in mset1, i.e. 3,
c          4,5, ... etc. Save those
c          with mismatches .ge. ndiff
c          in matrix mset2 starting at
c          position 2.
c          Next transfer contents
c          of mset2 into mset1 and
c          start
c          comparisons again this time
c          starting with subunit 3.
c          Continue until all subunits
c          undergo the comparisons.
c
c
c      npass=0
c
c
c      continue
1700      kk=npass+2
c      npass=npass+1
c

```

```

C      do 1500 m=npass+2,jj
C          n=0
C          do 1600 j=1,nsub
C              if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
2              . mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
2              mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
C                  n=n+1
C              endif
1600          continue
C              if(n.ge.ndiff) then
C                  kk=kk+1
C                  do 1625 i=1,nsub
1625                      mset2(kk,i)=mset1(m,i)
C              endif
1500          continue
C
C                                  kk is the number of subunits
C                                  stored in mset2
C
C                                  Transfer contents of mset2
C                                  into mset1 for next pass.
C
C
C          do 2000 k=1, kk
C              do 2000 m=1,nsub
2000                  mset1(k,m)=mset2(k,m)
C          if(kk.lt.jj) then
C              jj=kk
C              goto 1700
C          endif
C
C
C          nset=nset+1
C          write(1,7009)
7009          format(/)
C          do 7008 k=1, kk
7008              write(1,7010) (mset1(k,m),m=1,nsub)
7010          format(4i1)
C          write(*,*)
C          write(*,120) kk,nset
120          format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000          continue
C          close(1)
C
C
C      end
C
C          *****
C          *****

```

## APPENDIX Ib

Exemplary computer program for generating  
minimally cross hybridizing sets  
(single stranded tag/single stranded tag complement)

```

Program tagN
c
c
c      Program tagN generates minimally cross-hybridizing
c      sets of subunits given i) N--subunit length, and ii)
c      an initial subunit sequence. tagN assumes that only
c      3 of the four natural nucleotides are used in the tags.
c
c      character*1 sub1(20)
c      integer*2 mset(10000,20), nbase(20)
c
c
c      write(*,*) 'ENTER SUBUNIT LENGTH'
c      read(*,100) nsub
100  format(i2)
c
c      write(*,*) 'ENTER SUBUNIT SEQUENCE'
c      read(*,110) (sub1(k),k=1,nsub)
110  format(20a1)
c
c
c      ndiff=10
c
c      Let a=1 c=2 g=3 & t=4
c
c
c      do 800 kk=1,nsub
c      if(sub1(kk).eq.'a') then
c      mset(1, kk)=1
c      endif
c      if(sub1(kk).eq.'c') then
c      mset(1, kk)=2
c      endif
c      if(sub1(kk).eq.'g') then
c      mset(1, kk)=3
c      endif
c      if(sub1(kk).eq.'t') then
c      mset(1, kk)=4
c      endif
800  continue
c
c
c      Generate set of subunits differing from
c      sub1 by at least ndiff nucleotides.
c
c      jj=1
c
c      do 1000 k1=1,3

```

```

do 1000 k2=1,3
do 1000 k3=1,3
do 1000 k4=1,3
do 1000 k5=1,3
do 1000 k6=1,3
do 1000 k7=1,3
do 1000 k8=1,3
do 1000 k9=1,3
do 1000 k10=1,3
do 1000 k11=1,3
do 1000 k12=1,3
do 1000 k13=1,3
do 1000 k14=1,3
do 1000 k15=1,3
do 1000 k16=1,3
do 1000 k17=1,3
do 1000 k18=1,3
do 1000 k19=1,3
do 1000 k20=1,3

c
c

nbase(1)=k1
nbase(2)=k2
nbase(3)=k3
nbase(4)=k4
nbase(5)=k5
nbase(6)=k6
nbase(7)=k7
nbase(8)=k8
nbase(9)=k9
nbase(10)=k10
nbase(11)=k11
nbase(12)=k12
nbase(13)=k13
nbase(14)=k14
nbase(15)=k15
nbase(16)=k16
nbase(17)=k17
nbase(18)=k18
nbase(19)=k19
nbase(20)=k20

c
c

do 1250 nn=1,jj
n=0
do 1200 j=1,nsup
1   if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
2   mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
3   mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
mset(nn,j).eq.4 .and. nbase(j).ne.4) then
n=n+1
endif
1200 continue
c
c

if(n.lt.ndiff) then
goto 1000
endif
1250 continue
c
c

jj=jj+1
write(*,130) (nbase(i),i=1,nsup),jj
do 1100 i=1,nsup

```

```

                                mset(jj,i)=nbase(i)
1100                            continue
C
C
1000    continue
C
C
                                write(*,*)
130      format(10x,20(1x,i1),5x,i5)
                                write(*,*)
                                write(*,120) jj
120      format(1x,'Number of words=',i5)
C
C
                                end
C
C
                                *****
C
                                *****
C
```

**APPENDIX Ic**  
**Exemplary computer program for generating**  
**minimally cross hybridizing sets**  
**(double stranded tag/single stranded tag complement)**

```

Program 3tagN
C
C
C      Program 3tagN generates minimally cross-hybridizing
C      sets of duplex subunits given i) N--subunit length,
C      and ii) an initial homopurine sequence.
C
C      character*1 sub1(20)
C      integer*2 mset(10000,20), nbase(20)
C
C
C      write(*,*) 'ENTER SUBUNIT LENGTH'
C      read(*,100) nsub
100  format(i2)
C
C
C      write(*,*) 'ENTER SUBUNIT SEQUENCE a & g only'
C      read(*,110) (sub1(k),k=1,nsub)
110  format(20a1)
C
C      ndiff=10
C
C      Let a=1 and g=2
C
C      do 800 kk=1,nsub
C      if(sub1(kk).eq.'a') then
C      mset(1, kk)=1
C      endif
C      if(sub1(kk).eq.'g') then
C      mset(1, kk)=2
C      endif
800  continue
C
C      jj=1
C
C      do 1000 k1=1,3
C      do 1000 k2=1,3
C      do 1000 k3=1,3
C      do 1000 k4=1,3
C      do 1000 k5=1,3
C      do 1000 k6=1,3
C      do 1000 k7=1,3
C      do 1000 k8=1,3
C      do 1000 k9=1,3
C      do 1000 k10=1,3
C      do 1000 k11=1,3
C      do 1000 k12=1,3
C      do 1000 k13=1,3
C      do 1000 k14=1,3
C      do 1000 k15=1,3
C      do 1000 k16=1,3
C      do 1000 k17=1,3
C      do 1000 k18=1,3

```

```

do 1000 k19=1,3
do 1000 k20=1,3

c
nbase(1)=k1
nbase(2)=k2
nbase(3)=k3
nbase(4)=k4
nbase(5)=k5
nbase(6)=k6
nbase(7)=k7
nbase(8)=k8
nbase(9)=k9
nbase(10)=k10
nbase(11)=k11
nbase(12)=k12
nbase(13)=k13
nbase(14)=k14
nbase(15)=k15
nbase(16)=k16
nbase(17)=k17
nbase(18)=k18
nbase(19)=k19
nbase(20)=k20

c
do 1250 nn=1,jj
c
n=0
do 1200 j=1,nsup
  if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
1    mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
2    mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
3    mset(nn,j).eq.4 .and. nbase(j).ne.4) then
    n=n+1
  endif
1200  continue
c
  if(n.lt.ndiff) then
    goto 1000
  endif
1250  continue
c
  jj=jj+1
  write(*,130) (nbase(i),i=1,nsup),jj
  do 1100 i=1,nsup
    mset(jj,i)=nbase(i)
1100  continue
c
1000  continue
c
  write(*,*)
130  format(10x,20(1x,i1),5x,i5)
  write(*,*)
  write(*,120) jj
120  format(1x,'Number of words=',i5)
c
c
end

```

## SEQUENCE LISTING

## (1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

## (iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.  
(B) STREET: 3832 Bay Center Place  
(C) CITY: Hayward  
(D) STATE: California  
(E) COUNTRY: USA  
(F) ZIP: 94545

## (v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: 3.5 inch diskette  
(B) COMPUTER: IBM compatible  
(C) OPERATING SYSTEM: Windows 3.1  
(D) SOFTWARE: Microsoft Word 5.1

## (vi) CURRENT APPLICATION DATA:

(A) APPLICATION NUMBER:  
(B) FILING DATE:  
(C) CLASSIFICATION:

## (vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US96/09513  
(B) FILING DATE: 06-JUN-96

## (viii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: PCT/US95/12791  
(B) FILING DATE: 12-OCT-95

## (ix) ATTORNEY/AGENT INFORMATION:

(A) NAME: Stephen C. Macevicz  
(B) REGISTRATION NUMBER: 30,285  
(C) REFERENCE/DOCKET NUMBER: 813wo

## (x) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (510) 670-9365  
(B) TELEFAX: (510) 670-9302

## (2) INFORMATION FOR SEQ ID NO: 1:

## (i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 11 nucleotides  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A

11

(2) INFORMATION FOR SEQ ID NO: 2:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 11 nucleotides
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N

11

(2) INFORMATION FOR SEQ ID NO: 3:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 38 nucleotides
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA

38

(2) INFORMATION FOR SEQ ID NO: 4:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 20 nucleotides
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: double
  - (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG

20

(2) INFORMATION FOR SEQ ID NO: 5:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 20 nucleotides
  - (B) TYPE: nucleic acid

(C) STRANDEDNESS: double  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCACT CAGCGTCGAT

20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 20 nucleotides  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC

16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 62 nucleotides  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: double  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC  
TCTTCCTCTT CC

50

62

I claim:

1. A method of determining the toxicity of a compound, the method comprising the steps of:
  - 5 administering the compound to a test organism;  
extracting a population of mRNA molecules from each of one or more tissues of the test organism;  
forming a separate population of cDNA molecules from each population of mRNA molecules from the one or more tissues such that each cDNA molecule of a  
10 separate population has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;  
separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached;  
15 sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;  
determining the nucleotide sequence of a portion of each of the sorted cDNA  
20 molecules of each separate population to form a frequency distribution of expressed genes for each of the one or more tissues; and  
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.
- 25 2. The method of claim 1 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.
3. The method of claim 2 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in  
30 length and each subunit being selected from the same minimally cross-hybridizing set.
4. The method of claim 3 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation  
35 of unloaded microparticles.
5. The method of claim 4 further including a step of separating said loaded microparticles from said unloaded microparticles.

6. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.
- 5
7. The method of claim 6 wherein said number of loaded microparticles is at least 100,000.
8. The method of claim 7 wherein said number of loaded microparticles is at least 500,000.
- 10
9. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.
- 15
10. The method of claim 4 wherein said test organism is a mammalian tissue culture.
- 20
11. The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.
12. The method of claim 4 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
- 25
13. The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
- 30
14. A method of identifying genes which are differentially expressed in a selected tissue of a test animal after treatment with a compound, the method comprising the steps of:
- 35
- administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the test animal;

forming a population of cDNA molecules from the population of mRNA molecules such that each cDNA molecule has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by comparing the frequencing distribution of expressed genes of the selected tissue of the test animal with a frequency distribution of expressed genes of the selected tissue of a control animal.

15. The method of claim 14 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

16. The method of claim 15 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-hybridizing set.

17. The method of claim 16 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation of unloaded microparticles.

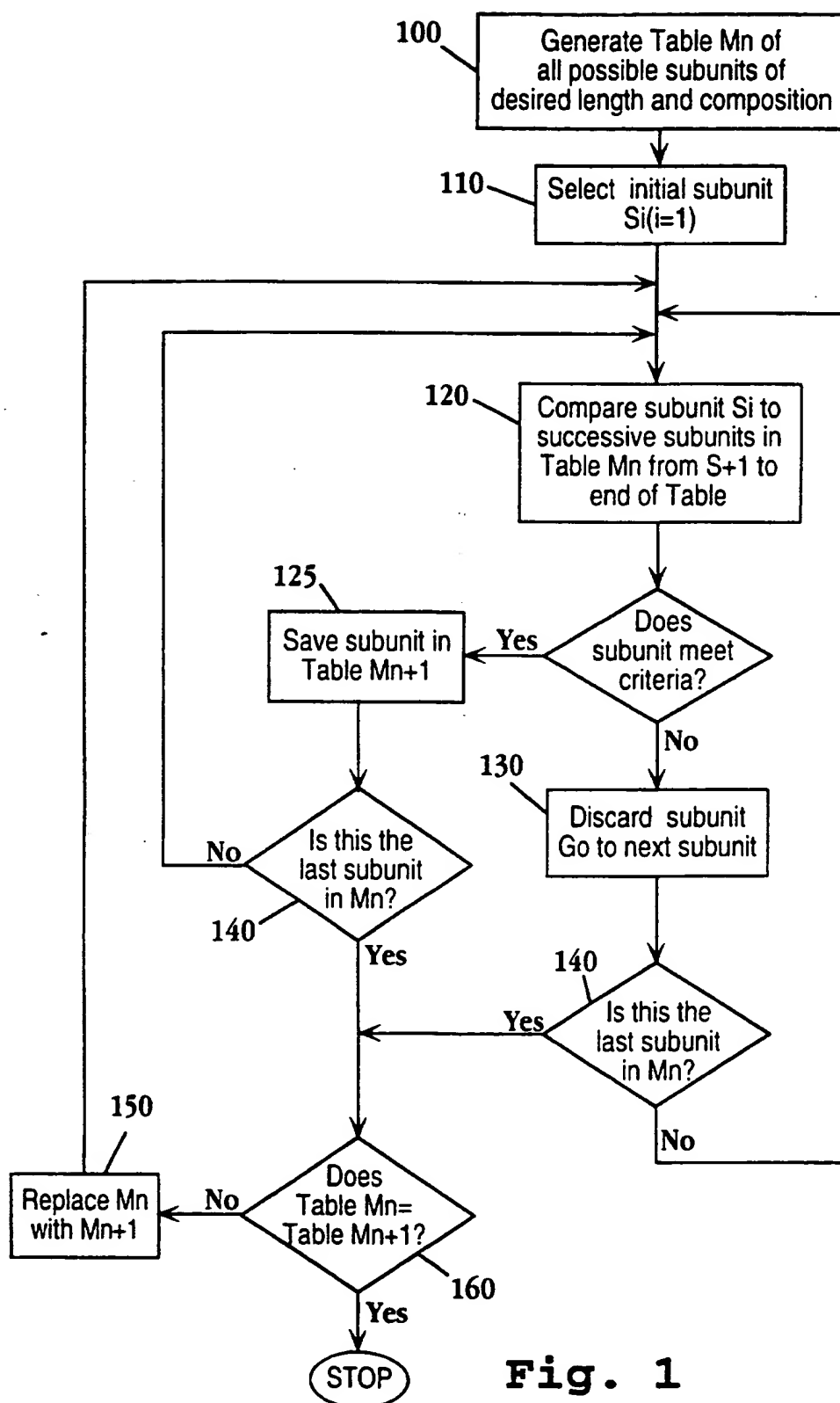
18. The method of claim 17 further including a step of separating said loaded microparticles from said unloaded microparticles.

19. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

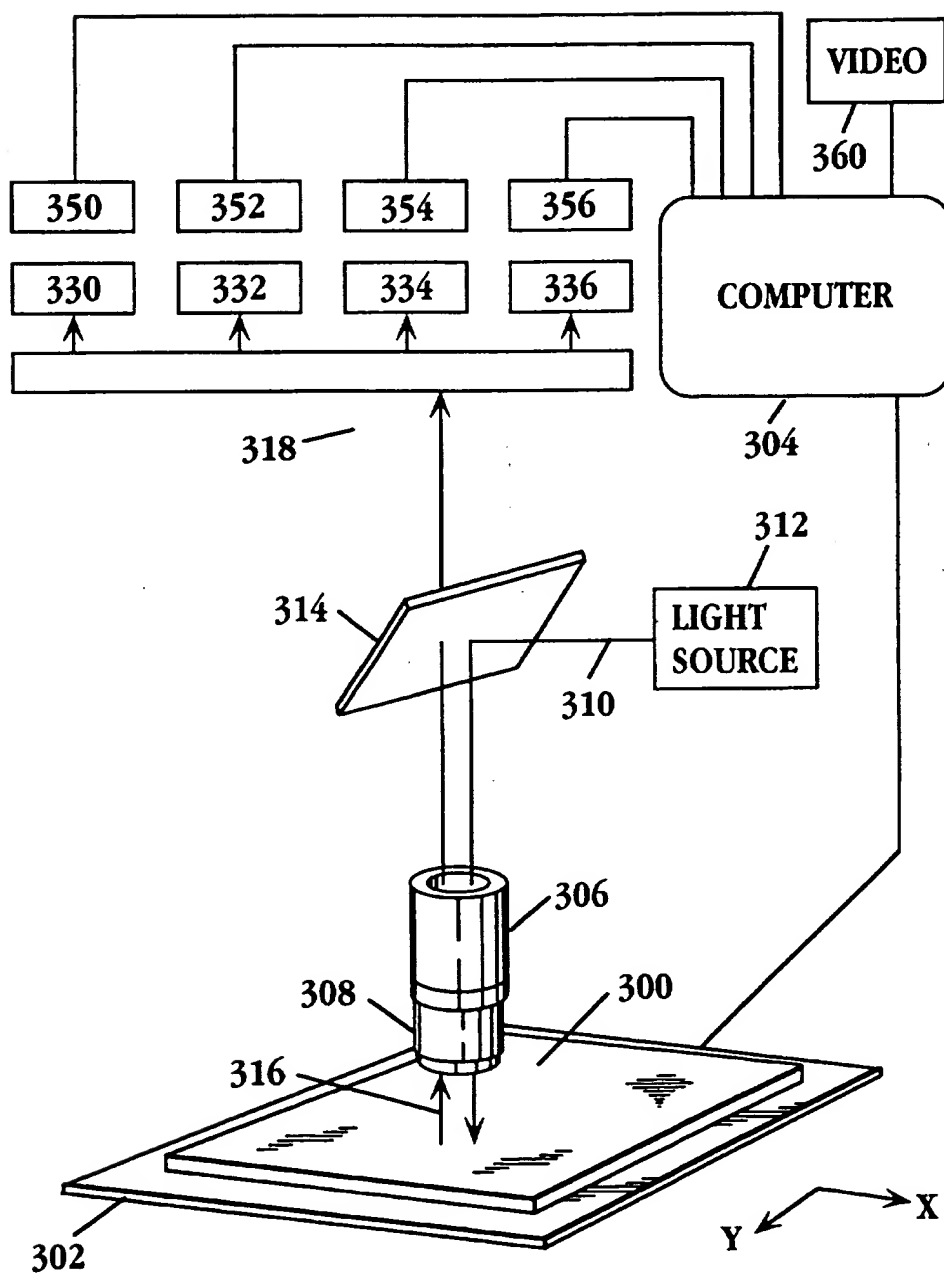
20. The method of claim 19 wherein said number of loaded microparticles is at least 100,000.
21. The method of claim 20 wherein said number of loaded microparticles is at least 500,000.
22. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.
23. The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
24. The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
25. A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:  
administering the compound to a test organism;  
extracting a population of mRNA molecules from each of one or more tissues of the test organism and forming a population of cDNA molecules for each of the one or more tissues;  
determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more tissues; and  
correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.
26. The use of claim 25 wherein said test organism is a mammalian tissue culture.
27. The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28. The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
- 5 29. The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancreas urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.
- 10 30. A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:
- 15 administering the compound to the test organism;
- extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;
- determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency distribution of expressed genes;
- 20 identifying genes expressed in response to administering the compound by comparing the frequency distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and
- 25 determining whether the genes expressed in response to administering the compound are correlated with toxicity of the compound in the test organism.

1/2

**Fig. 1**

2/2

**Fig. 2**

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/16342

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : C12Q 1/68; C07H 21/04

US CL : 435/6; 536/24.3

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH

search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?,

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096.	1-30
A	BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383.	1-30
A	MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274.	1-30

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E* earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G*	document member of the same patent family
*O* document referring to an oral disclosure, use, exhibition or other means		
*P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

27 JANUARY 1997

Date of mailing of the international search report

19 FEB 1997

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

SCOTT D. PRIEBE

Telephone No. (703) 308-0196

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/US96/16342

**C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4.	1-30

FOCUS - 17 of 19 DOCUMENTS

Copyright 1997 PR Newswire Association, Inc.  
PR Newswire

August 11, 1997, Monday

**SECTION:** Financial News

**DISTRIBUTION:** TO BUSINESS AND MEDICAL EDITORS

**LENGTH:** 478 words

**HEADLINE:** Eli Lilly & Co. and Acacia Biosciences Enter Into Research Collaboration;  
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

**DATELINE:** RICHMOND, Calif., Aug. 11

**BODY:**

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.

**SOURCE** Acacia Biosciences

**CONTACT:** Bruce Cohen, President and CEO of Acacia Biosciences, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

**LOAD-DATE:** August 12, 1997

**The Bioreactor Market: Steady Growth Expected**

The worldwide market for all bioreactors was valued at \$273 million for 1997, and is expected to be worth \$380 million by 2002.

Types: anaerobic, good

V. 17  
 C. 01  
 T1: GENETIC ENGINEERING NEWS

W1 GE281N  
 NO. 16  
 SEQ: G04575000  
 1997

09/25/97

# GENETIC ENGINEERING NEWS

BIOTECHNOLOGY • BIOPROCESS • BIORESEARCH • TECHNOLOGY TRANSFER

<b>Contents</b>	
European Biotech Standards Moving	4
Biotech Companies: Separation Packages	13
Trends in Biotechnology Development	14
Q/QA for Biotech Firms	16
Advances in Electroporation	19
New Products	21
New GEN Column: Drug Discovery, Assay Miniaturization	27
Corporate Profile: Pangen Systems	28
Corporate Market	29
European Roundup	30
With Stock Offers	31
Company Agreements	32
Inside Industry	33
China Trade Update	37
Not So Good	40
People	41
Calendar	41
Marketplace	42

## Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

**P**harmagene, the Royston, U.K.-based biopharmaceutical company specializing in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies

SEE PHARMAGENE, P. 9

## Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

**P**erkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Frammingham, MA) via a \$360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."

A driving force behind the merger is the vast amount of genet-



Perkin-Elmer acquired PerSeptive Biosystems for \$360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

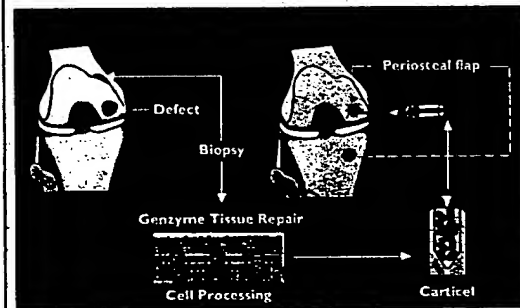
ic information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for

studying proteins and protein networks—a field known as proteomics (see GEN, September 1, 1997, p. 1).

PE officials, who claim that MALDI-TOF (Matrix Assisted

SEE ACQUISITION, P. 10

## FDA OKs Genzyme's Carticel Product for Damage to Knees



Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

**T**he FDA has approved a knee cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel (autologous cultured chondrocytes) is the first product to be licensed under the FDA's pro-

SEE GENZYME, P. 6

## Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

**A**ccacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Accacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Accacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout

for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

### Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

SEE TARGET, P. 18

## Sticky Ends

Avigen received two grants from the NIH & University of California for research on gene therapy for treatment of cancer & HIV infections...MRL Pharmaceutical Services, of Reston, VA, launched the TSN Bug Finder, which is able to locate & retrieve client-specified microorganisms in real-time...GenSis Sicor, Inc. will move its corporate staff from San Diego to Irvine, CA, by end of year...

FDA accepted NDA from Sepacor for levalbuterol HCl inhalation solution...An \$11.7M mezzanine financing has been closed by Activated Cell Therapy, which changed its name to Dendreon Corporation...Astra AB will build major research facility in Waltham, MA, and is also relocating Astra Arcus research facility from Rochester to Boston area...Prolifix Ltd. team used a small peptide to inhibit the E2F protein complex and induced

apoptosis in mammalian tumor cells...Vertex Pharmaceuticals, Inc. and Alpha Therapeutic Corp. ended an agreement to develop VX-366 for treatment of inherited hemoglobin disorders...NaviCyte received Phase I SBIR grant for up to \$100,000 from NIH for development of prototype of its NaviFlow technology for high-throughput screening...Covance Inc. will invest \$21 million in expansion and renovation of its facility in Indianapolis, IN.



# Target

from page 1

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

Spinning off technology developed at the University of North Carolina (Chapel Hill), Cytogen Corp. (Princeton, NJ) formed its wholly owned subsidiary AxCell Biosciences earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a

range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (GDL) to map functional protein interactions and quantitate their affinity. The company's inter-functional proteomic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

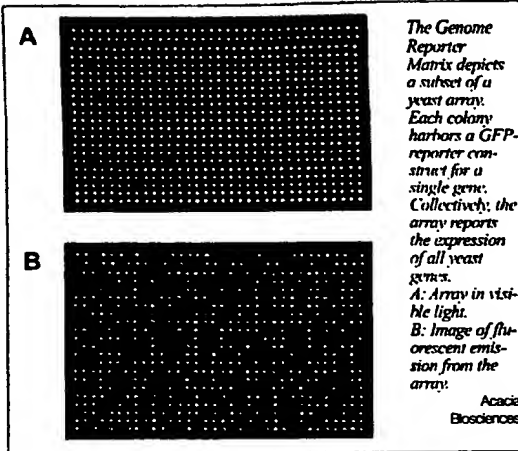
## Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38

signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification

programs focus on the NF- $\kappa$ B pathway, estrogen-related genes and central/peripheral nervous system genes.

Regulating cytokine production in immune and inflammatory disorders,



**The Genome Reporter Matrix** depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes.  
A: Array in visible light.  
B: Image of fluorescent emission from the array.

Acacia Biosciences

and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with Tanabe Selyaku (Osaka, Japan). Signal has partnered with Organon/Alzo Nobel (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with Roche Bioscience (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis' (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of *Drosophila* and *Caenorhabditis elegans*. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with SmithKline (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in Axiom Biotechnologies (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, gene/Networks (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiological traits.

## Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at Ribozyme Pharmaceuticals, Inc. (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

Correlation of the gene expression inhibition with phenotype can

SEE TARGET, P. 38

## A strong chemical combination to help you grow. And flourish.

Three hundred million dollars and ten years of hard work. That's what it costs to bring your biotechnology-derived therapeutic to the marketplace.

Which means, no room for error.

Which means, in turn, you'd be wise to tap into the combined capabilities of Mallinckrodt and J.T.Baker: dual sources, trusted names for your chemical raw materials.

Two separate GMP-produced brands offering the control of a single quality system and the convenience of a single audit process.

We offer comprehensive product lines including USP salts, bioreagents, high purity solvents and chromatography products in Beaker to Bulk™ packaging for easy scale-up.

Call 1-800-582-2537, or access our website at <http://www.mallinckrodt.com>. For dual chemical sources dedicated to helping you grow. Flourish. Succeed!

MALLINCKRODT



## Target

from page 15

suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with Schering AG (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with Chiron Corp. (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, Sequitur, Inc. (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target, Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

### Mining Pathogen Genomes

Companies such as Human Genome Sciences (HGS; Rockville, MD), Incyte (Palo Alto, CA),

*ArCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isolated using a combination structural/functional approach that uses both peptide phase display Genetic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.*

Millennium Pharmaceuticals Inc. (Cambridge, MA) and Genome Therapeutics (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with Hoffmann-La Roche (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identifying

60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF- $\beta$ ) signaling. The company also received U.S. patent coverage for the tub genes, and for the gene that encodes the protein metastatin, which appears to suppress metastasis in malignant melanoma.

## Pangea

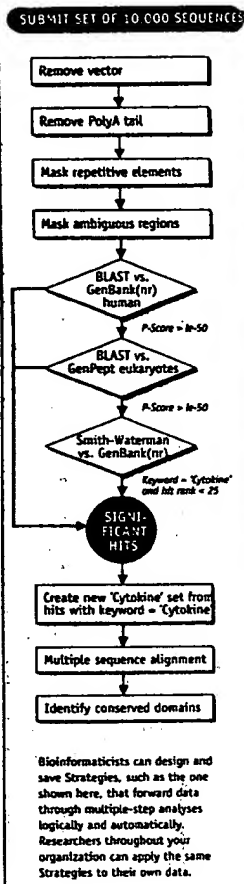
from page 28

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with



commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator provides a friendly user interface from PC, Macintosh, and UNIX workstations.

In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson.

## Europe

from page 30

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clotier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hurler's syndrome (1), cervical cancer (1), glioblastoma

breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC third annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

### Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted polymethylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the



## NEW HIGH SPECIFIC ACTIVITY MICROBIAL ALKALINE PHOSPHATASE from Biocatalysts

Biocatalysts Limited, the British speciality enzyme company, has developed a completely new type of alkaline phosphatase with many advantages over the types most commonly used.

It is of microbial origin with a high specific activity (unlike that from *E. coli*) and with higher temperature and storage stability compared to that from calf intestine.

This is the first of several new generation diagnostic enzymes being developed by Biocatalysts Limited with greatly improved stability.

- Non-animal source, no risk of BSE or animal virus contamination
- Higher temperature stability than calf intestine
- Much higher specific activity than from *E. coli*
- Very high storage stability even in the absence of glycerol

For further details on alkaline phosphatase and our other diagnostic enzymes contact us direct at the address below or within North America contact our US Distributor Kaitron-Petibone 'phone: 630 350 1116 or fax: 630-350-1606

Biocatalysts Limited  
Treforest Industrial Estate Pontypridd Wales UK CF37 5UD  
Tel: +44 (0)1443 843712 Fax: +44 (0)1443 841214  
e-mail: kelly@biocatalysts.com.



- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch et al., *EMBO J.* 13, 3822 (1994); M. T. Madiredi et al., *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari et al., *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Megannathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa et al., *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho et al., *Cell* 77, 869 (1994).
38. W. Hendriks et al., *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

## Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown\*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

*Saccharomyces cerevisiae* is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity, measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

\*To whom correspondence should be addressed. E-mail: pbrown@crgm.stanford.edu

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl-coenzyme A (CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

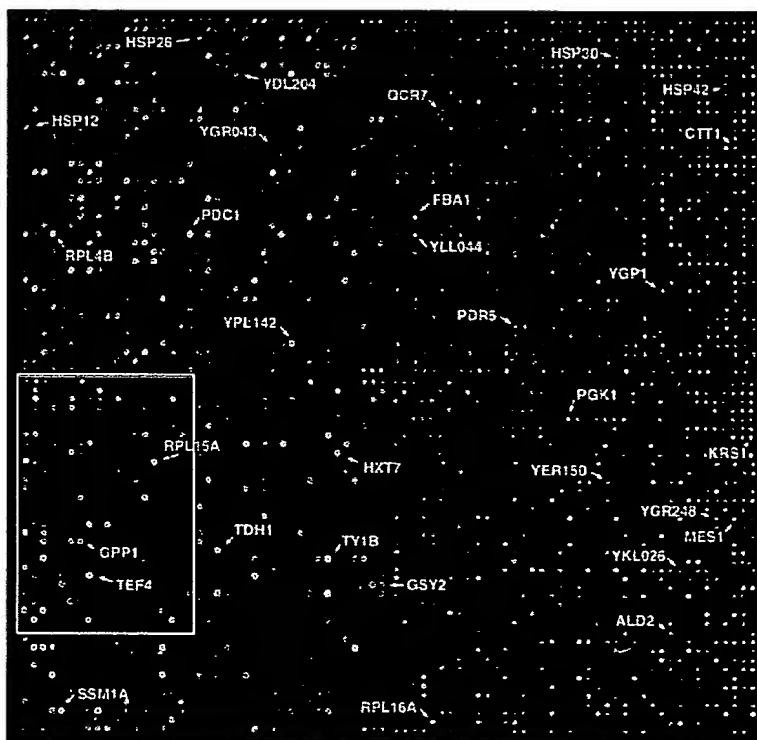
Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception



**Fig. 1.** Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of  $<5 \times 10^6$  cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of  $\sim 2 \times 10^8$  cells/ml, with a glucose level of  $<0.2$  g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGG (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome *c*-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome *c*-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS<sub>rp</sub>) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the “master regulator” of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of

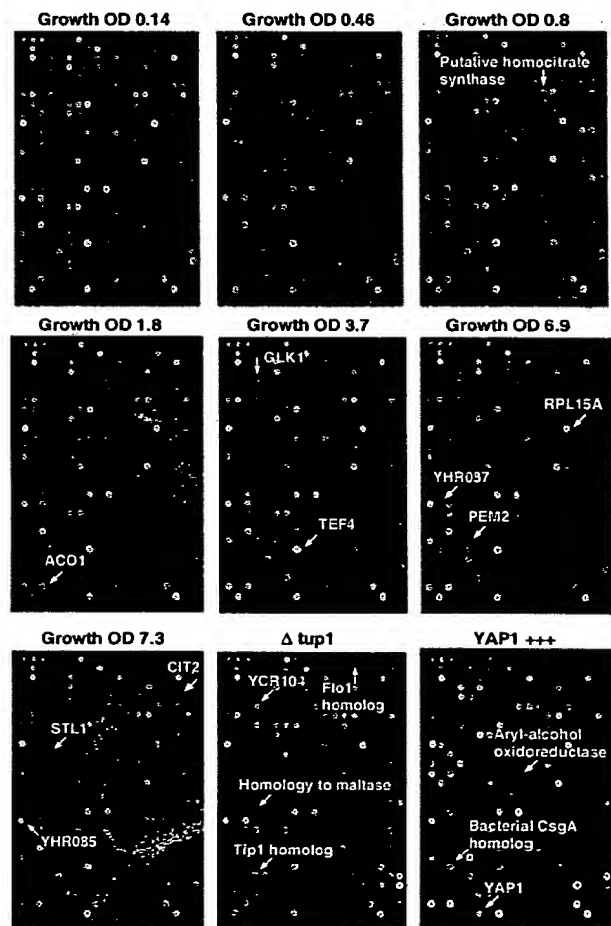
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

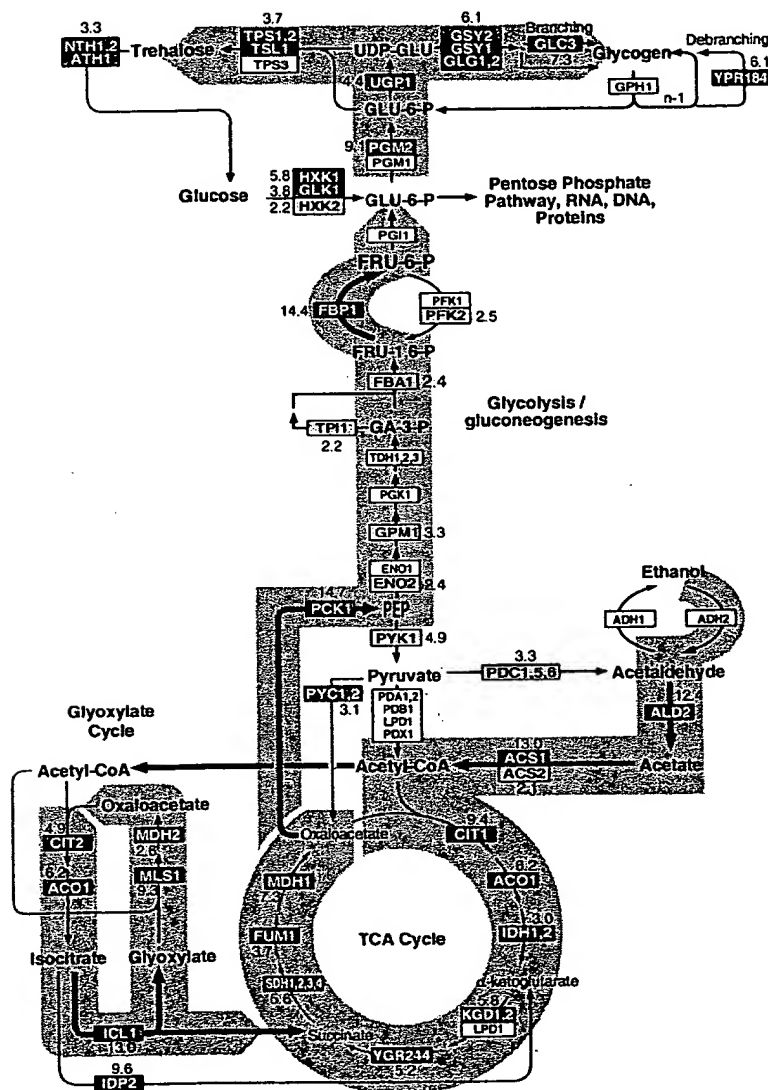
The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

**Fig. 2.** The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as *Tipl* and *Tir1/Srp1* which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*



www.sciencemag.org • SCIENCE • VOL. 278 • 24 OCTOBER 1997

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT  $\alpha$  strain in which *MFA1* and *MFA2*, the genes encoding the  $\alpha$ -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1 $\Delta$*  strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MATA strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

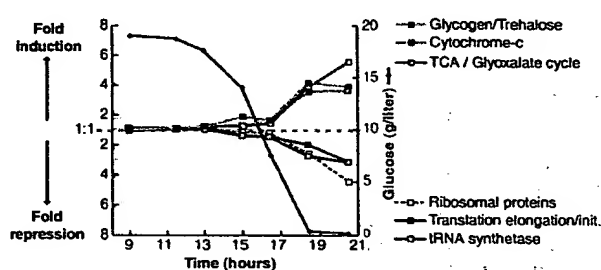
Of the 17 genes whose mRNA levels increased by more than threefold when

*YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

**Fig. 4.** Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.



**Table 1.** Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C		<i>YAP1</i>	Putative aryl-alcohol reductase	12.9
YKL071W	162-222 (5 sites)		Similarity to bacterial <i>csgA</i> protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W		<i>ATR1</i>	Putative aryl-alcohol reductase	6.5
YML116W	409		Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	<i>OYE3</i>	NADPH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

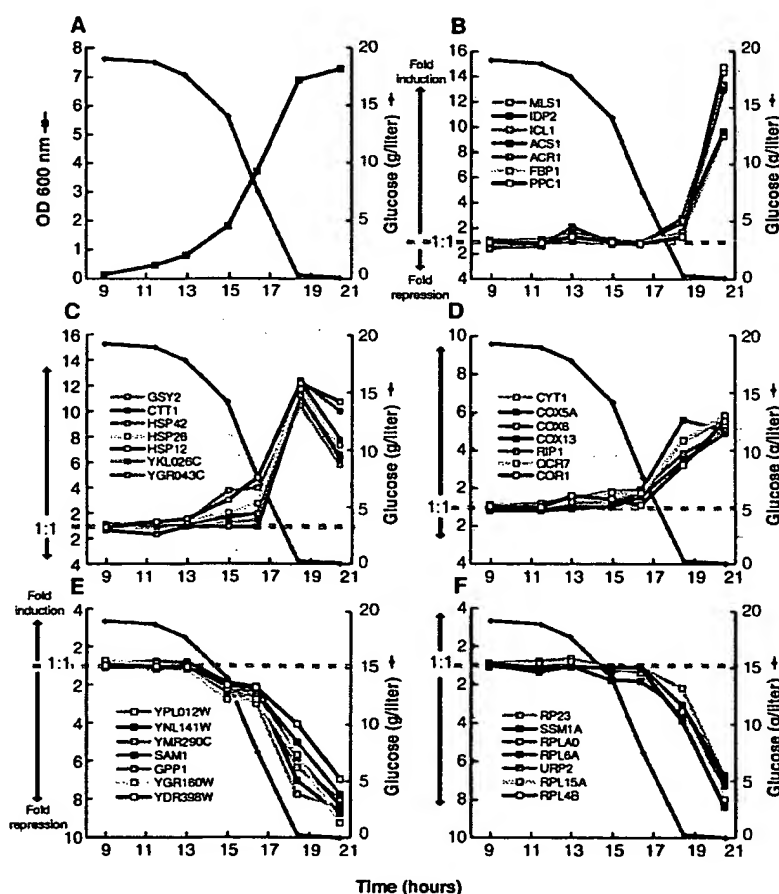
works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

## REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- $\mu$ l PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3 $\times$  standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarray are available at [cmgm.stanford.edu/pbrown](http://cmgm.stanford.edu/pbrown). After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-



**Fig. 5.** Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contains STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at  $-95^{\circ}\text{C}$ . The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at  $30^{\circ}\text{C}$  with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at  $-80^{\circ}\text{C}$ .
  11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25  $\mu\text{g}$  of polyadenylated [poly(A) $^{+}$ ] RNA, primed by a dT(16) oligomer. This mixture was heated to  $70^{\circ}\text{C}$  for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500  $\mu\text{M}$  for dATP, dCTP, and dGTP and 200  $\mu\text{M}$  for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100  $\mu\text{M}$ . The reaction was then incubated at  $42^{\circ}\text{C}$  for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with 470  $\mu\text{l}$  of 10 mM Tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to  $\sim 5 \mu\text{l}$  using Centricon-30 microconcentrators (Amicon).
  12. Purified, labeled cDNA was resuspended in 11  $\mu\text{l}$  of 3.5 $\times$  SSC containing 10  $\mu\text{g}$  poly(dA) and 0.3  $\mu\text{l}$  of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for  $\sim 8$  to 12 hours in a water bath at  $62^{\circ}\text{C}$ . Before scanning, slides were washed in 2 $\times$  SSC, 0.2% SDS for 5 min, and then 0.05 $\times$  SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
  13. The complete data set is available on the Internet at [cmgm.stanford.edu/pbrown/explore/index.html](http://cmgm.stanford.edu/pbrown/explore/index.html)
  14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
  15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database ([genome-www.stanford.edu](http://genome-www.stanford.edu)), Yeast Protein Database ([quest7.proteome.com](http://quest7.proteome.com)), and Munich Information Centre for Protein Sequences ([speedy.mips.biochem.mpg.de/mips/yeast/index.htm](http://speedy.mips.biochem.mpg.de/mips/yeast/index.htm)).
  16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3613 (1994).
  17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
  18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* 268, 12116 (1993).
  19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Genet.* 242, 727 (1994).
  20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
  21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
  22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
  23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
  24. J. L. Parrou, M. A. Testa, J. Francois, *Microbiology* 143, 1891 (1997).
  25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
  26. S. L. Forsburg and L. Guarante, *Genes Dev.* 3, 1166 (1989).
  27. J. T. Olesen and L. Guarante, *ibid.* 4, 1714 (1990).
  28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Deverish, *Mol. Microbiol.* 13, 119 (1994).
  29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B-C, G, or T; N-G, A, T, or C; R-A or G; and Y-C or T.
  30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
  31. D. Shore, *Trends Genet.* 10, 408 (1994).
  32. R. J. Planta and H. A. Raue, *ibid.* 4, 64 (1988).
  33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.
  34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
  35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
  36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PFYK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *TCT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
  37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).
  38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXX1/HXX2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) (20), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
  39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
  40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
  41. Differences in mRNA levels between the *tup1 $\Delta$*  and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1 $\Delta$*  strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
  42. The *tup1 $\Delta$*  mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
  43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
  44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
  45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
  46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
  47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
  48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
  49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at [cmgm.stanford.edu/pbrown](http://cmgm.stanford.edu/pbrown). Images were scanned at a resolution of 20  $\mu\text{m}$  per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
  50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
  51. We thank H. Bennett, P. Spelman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop PCT, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 3, 2004.

By: \_\_\_\_\_ Printed: Lisa McDill

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of: Sanjanwala et al.

Title: ENZYMES

Serial No.: 10/467,903

Filing Date: Herewith

Examiner: To Be Assigned

Group Art Unit: To Be Assigned

Mail Stop PCT

Commissioner for Patents

P.O. Box 1450

Alexandria, VA 22313-1450

**TRANSMITTAL FEE SHEET**

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Petition for Extension of Time under 37 CFR § 1.136 (1 pg.);
3. Response to Notice to File Missing Requirements under 35 USC § 371 (2 pp.);
4. Copy of Notice to File Missing Requirements under 35 USC § 371 dated December 3, 2003 (2 pp.);
5. Executed Declaration and Power of Attorney (66 pp., signed in counter-part);
6. Petition Under 1.147 (a) (4 pp.);
7. Declaration of Nancy Ramos (3 pp.), with Exhibits A-C; and
8. Declaration of Katherine Stofer (2 pp.).

The fee has been calculated as shown below:

<input checked="" type="checkbox"/> Fee for filing a Petition for Extension of Time Under 37 CFR 1.17(a) -	<u>1</u> Mo(s):	\$ 110.00
<input checked="" type="checkbox"/> Fee for filing late Oath or Declaration under 37 CFR § 1.492(e):		\$ 130.00
<input checked="" type="checkbox"/> Please charge Deposit Account No. 09-0108 in the amount of :		\$ <u>240.0</u>

The Commissioner is hereby authorized to charge any additional fees required under 37 CFR 1.16 and 1.17, or credit overpayment to Deposit Account No. 09-0108. A duplicate copy of this sheet is enclosed.

Respectfully submitted,

INCYTE CORPORATION

Date: March 3, 2004

\_\_\_\_\_  
Richard C. Ekstrom

Reg. No. 37,027

Direct Dial Telephone: (650) 843-7352

Customer No.: 27904

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555 or Fax: (650) 845-4166

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop PCT, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 3, 2004.

By: \_\_\_\_\_ Printed: Lisa McDill

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of: Sanjanwala et al.

Title: ENZYMES

Serial No.: 10/467,903

Filing Date: Herewith

Examiner: To Be Assigned

Group Art Unit: To Be Assigned

Mail Stop PCT  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

**TRANSMITTAL FEE SHEET**

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Petition for Extension of Time under 37 CFR § 1.136 (1 pg.);
3. Response to Notice to File Missing Requirements under 35 USC § 371 (2 pp.);
4. Copy of Notice to File Missing Requirements under 35 USC § 371 dated December 3, 2003 (2 pp.);
5. Executed Declaration and Power of Attorney (66 pp., signed in counter-part);
6. Petition Under 1.147 (a) (4 pp.);
7. Declaration of Nancy Ramos (3 pp.), with Exhibits A-C; and
8. Declaration of Katherine Stofer (2 pp.).

The fee has been calculated as shown below:

<u>X</u>	Fee for filing a Petition for Extension of Time Under 37 CFR 1.17(a) -	<u>1</u> Mo(s):	\$ 110.00
<u>X</u>	Fee for filing late Oath or Declaration under 37 CFR § 1.492(e):		\$ 130.00
<u>X</u>	Please charge Deposit Account No. <b>09-0108</b> in the amount of :		\$ <u>240.0</u>

The Commissioner is hereby authorized to charge any additional fees required under 37 CFR 1.16 and 1.17, or credit overpayment to Deposit Account No. 09-0108. A duplicate copy of this sheet is enclosed.

Respectfully submitted,

INCYTE CORPORATION

Date: March 3, 2004

\_\_\_\_\_  
Richard C. Ekstrom  
Reg. No. 37,027  
Direct Dial Telephone: (650) 843-7352

Customer No.: 27904  
3160 Porter Drive  
Palo Alto, California 94304  
Phone: (650) 855-0555 or Fax: (650) 845-4166



Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Boo

Search  for

Limits

Preview/Index

History

Clipboard

Details

Show:

☐ 1: BAC05916. seven transmembra...[gi:21928657]

[BLink](#), [Domains](#), [Links](#)

LOCUS BAC05916 310 aa linear PRI 23-JUL-2002

DEFINITION seven transmembrane helix receptor [Homo sapiens].

ACCESSION BAC05916

VERSION BAC05916.1 GI:21928657

DBSOURCE accession AB065693.1

KEYWORDS

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

REFERENCE

1  
AUTHORS Suwa,M., Sato,T., Okouchi,I., Arita,M., Futami,K., Matsumoto,S., Tsutsumi,S., Aburatani,H., Asai,K. and Akiyama,Y.

TITLE Genome-wide discovery and analysis of human seven transmembrane helix receptor genes

JOURNAL Unpublished

REFERENCE 2 (residues 1 to 310)

AUTHORS Suwa,M.

TITLE Direct Submission

JOURNAL Submitted (11-JUL-2001) Makiko Suwa, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST); 2-41-6 Aomi Koto-ku, Tokyo 135-0064, Japan (E-mail:m-suwa@aist.go.jp, URL:http://www.cbrc.jp/, Tel:81-3-3599-8080, Fax:81-3-3599-8081)

COMMENT

This sequence is a seven transmembrane helix receptor candidate predicted from the whole human genome sequences using our automated system that contains programs of gene finding(GeneDecoder), sequence search, motif-domain assignment and transmembrane helix prediction.

And the sequence is submitted by the collaborative project between [Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)] and [Genome Science Division, Research Center for Advanced Science and Technology (RCAT), University of Tokyo].

FEATURES

source

Location/Qualifiers

1..310

/organism="Homo sapiens"

/isolate="CBRC7TM\_256"

/db\_xref="taxon:9606"

/chromosome="7"

Protein

1..310

/product="seven transmembrane helix receptor"

CDS

1..310

/coded\_by="AB065693.1:201..1133"

ORIGIN

```
1 megnktwitd itlprfqvvp aleillcglf safytltlhg ngvifgiicl dcklhtpmf
61 flshlaivdi syasnyvpkm ltnlmngest isffpcimqt flylafahve clilvmsyd
121 ryadichplr ynslmswrvv tvlavaswvf sllalvplv lilsfpfcgp heinhffcei
181 lsvlklacad twlnqviva acvfilvgpl clvlvsylri laailriqsg egrrkafstc
```

241 sshlcvgglf fgsaivtyma pksrhpeeqq kvlslfyslf npmlnpliys lrnaevkga1  
301 rralrkerlt

//

[Disclaimer](#) | [Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)

Feb 24 2004 16:01:25

Docket No.:PI-0360-USN  
USSN:10/467,903  
Exhibit A

Docket No.:PI-0360-USN  
USSN:10/467,903  
Exhibit B

Docket No.:PI-0360-USN  
USSN:10/467,903  
Exhibit C

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop PCT, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 3, 2004.

By: \_\_\_\_\_ Printed: Lisa McDill

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of: Sanjanwala et al.

Title: ENZYMES

Serial No.: 10/467,903

Filing Date: Herewith

Examiner: To Be Assigned

Group Art Unit: To Be Assigned

Mail Stop PCT  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

**TRANSMITTAL FEE SHEET**

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Petition for Extension of Time under 37 CFR § 1.136 (1 pg.);
3. Response to Notice to File Missing Requirements under 35 USC § 371 (2 pp.);
4. Copy of Notice to File Missing Requirements under 35 USC § 371 dated December 3, 2003 (2 pp.);
5. **Executed** Declaration and Power of Attorney (66 pp., signed in counter-part);
6. Petition Under 37 C.F.R. 1.47 (a) (4 pp.);
7. Declaration of Nancy Ramos (3 pp.), with Exhibits A-C; and
8. Declaration of Katherine Stofer (2 pp.).

The fee has been calculated as shown below:

<u>X</u> Fee for filing a Petition for Extension of Time Under 37 CFR 1.17(a) -	<u>1</u> Mo(s):	\$ 110.00
<u>X</u> Fee for filing late Oath or Declaration under 37 CFR § 1.492(e):		\$ 130.00
<u>X</u> Please charge Deposit Account No. <b>09-0108</b> in the amount of :		\$ <u>240.0</u>

The Commissioner is hereby authorized to charge any additional fees required under 37 CFR 1.16 and 1.17, or credit overpayment to Deposit Account No. 09-0108. **A duplicate copy of this sheet is enclosed.**

Respectfully submitted,

INCYTE CORPORATION

Date: March 3, 2004

\_\_\_\_\_  
Richard C. Ekstrom  
Reg. No. 37,027  
Direct Dial Telephone: (650) 843-7352

**Customer No.: 27904**  
3160 Porter Drive  
Palo Alto, California 94304  
Phone: (650) 855-0555 or Fax: (650) 845-4166

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop PCT, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 3, 2004.

By: \_\_\_\_\_ Printed: Lisa McDill

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of: Sanjanwala et al.

Title: ENZYMES

Serial No.: 10/467,903

Filing Date: Herewith

Examiner: To Be Assigned

Group Art Unit: To Be Assigned

Mail Stop PCT

Commissioner for Patents

P.O. Box 1450

Alexandria, VA 22313-1450

**TRANSMITTAL FEE SHEET**

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Petition for Extension of Time under 37 CFR § 1.136 (1 pg.);
3. Response to Notice to File Missing Requirements under 35 USC § 371 (2 pp.);
4. Copy of Notice to File Missing Requirements under 35 USC § 371 dated December 3, 2003 (2 pp.);
5. Executed Declaration and Power of Attorney (66 pp., signed in counter-part);
6. Petition Under 37 C.F.R. 1.47 (a) (4 pp.);
7. Declaration of Nancy Ramos (3 pp.), with Exhibits A-C; and
8. Declaration of Katherine Stofer (2 pp.).

The fee has been calculated as shown below:

<u>X</u> Fee for filing a Petition for Extension of Time Under 37 CFR 1.17(a) -	<u>1</u> Mo(s):	\$ 110.00
<u>X</u> Fee for filing late Oath or Declaration under 37 CFR § 1.492(e):		\$ 130.00
<u>X</u> Please charge Deposit Account No. 09-0108 in the amount of :		\$ <u>240.0</u>

The Commissioner is hereby authorized to charge any additional fees required under 37 CFR 1.16 and 1.17, or credit overpayment to Deposit Account No. 09-0108. A duplicate copy of this sheet is enclosed.

Respectfully submitted,

INCYTE CORPORATION

Date: March 3, 2004

\_\_\_\_\_  
Richard C. Ekstrom

Reg. No. 37,027

Direct Dial Telephone: (650) 843-7352

Customer No.: 27904

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555 or Fax: (650) 845-4166

- Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch et al., *EMBO J.* 13, 3822 (1994); M. T. Madireddi et al., *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
  36. P. M. Palosaari et al., *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

- Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa et al., *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho et al., *Cell* 77, 869 (1994).
38. W. Hendriks et al., *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tiford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

## Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown\*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

*Saccharomyces cerevisiae* is an especially

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

\* To whom correspondence should be addressed. E-mail: pbrown@crgm.stanford.edu

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl-coenzyme A (CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-bisphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

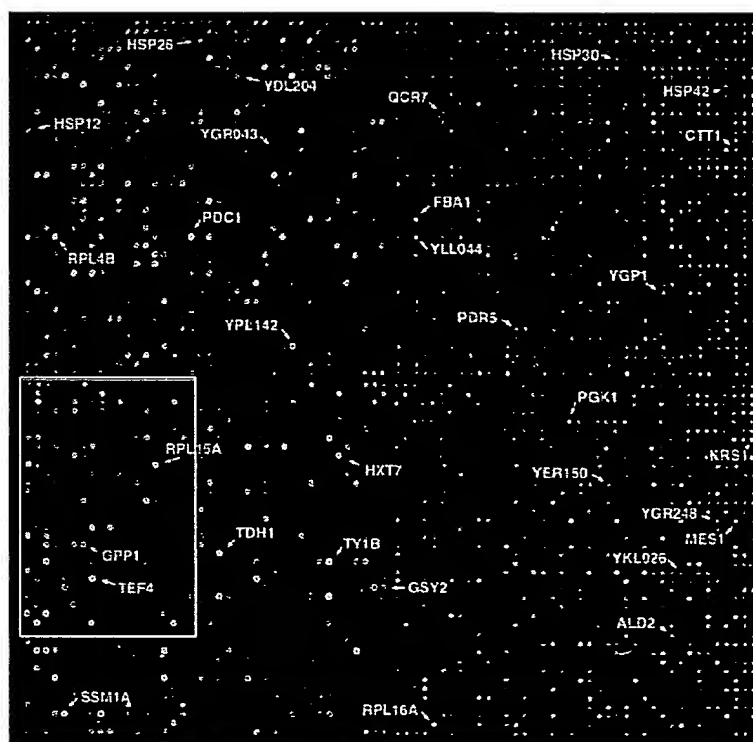
Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c-related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception



**Fig. 1.** Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of  $<5 \times 10^6$  cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of  $\sim 2 \times 10^8$  cells/ml, with a glucose level of  $<0.2$  g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP-labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP-labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGG (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome *c*-related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome *c*-related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS<sub>rp</sub>) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the “master regulator” of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of

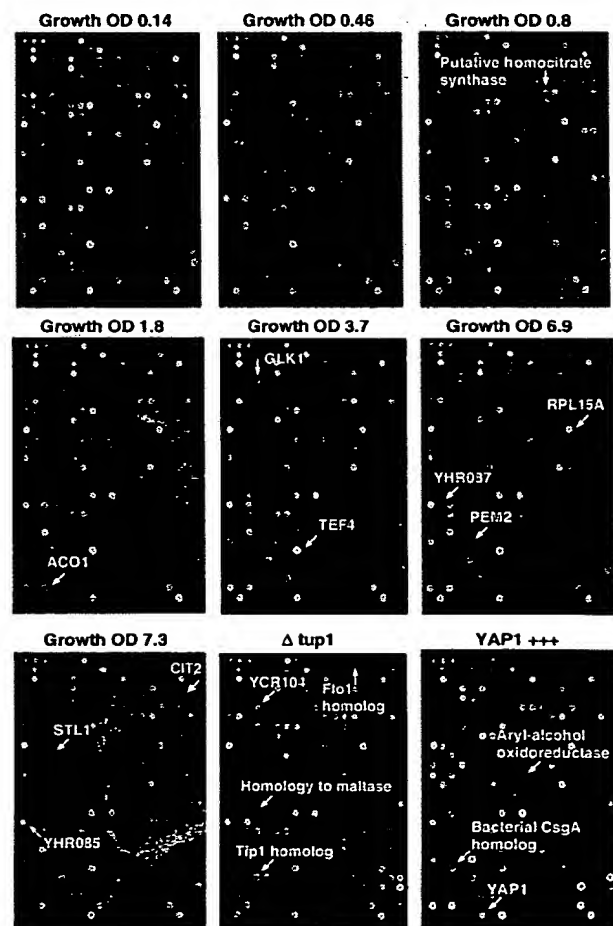
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

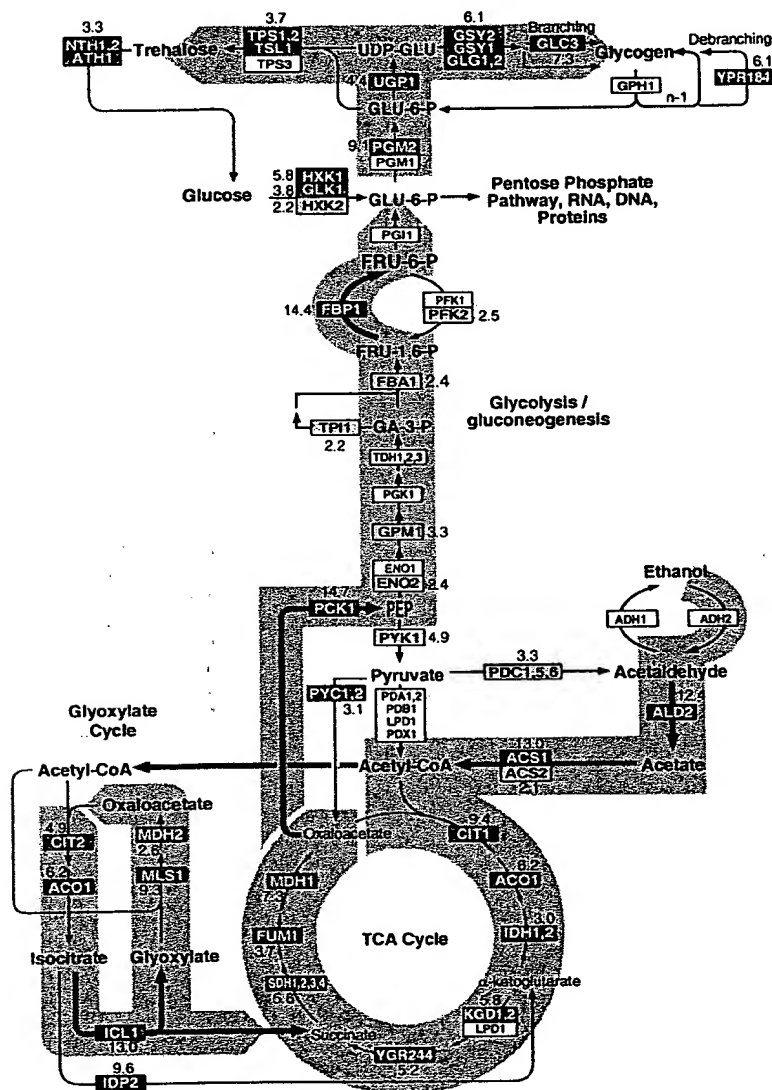
The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

**Fig. 2.** The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.



by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type-specific, and DNA-damage-inducible genes (40).



**Fig. 3.** Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolites, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included  $\alpha$ -glucosidases, the mating-type-specific genes *MFA1* and *MFA2*, and the DNA damage-inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT  $\alpha$  strain in which *MFA1* and *MFA2*, the genes encoding the  $\alpha$ -factor mating pheromone precursor, are normally repressed. In the isogenic *tup1 $\Delta$*  strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MATA strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GALI-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

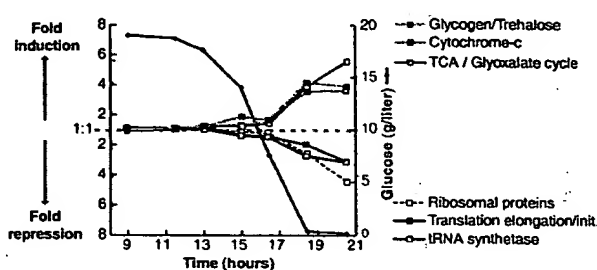
Of the 17 genes whose mRNA levels increased by more than threefold when

*YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing *Yap1*. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for *Yap1*-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon *Yap1* overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by *Yap1*. The absence of canonical *Yap1*-bind-

**Fig. 4.** Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondrial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.



**Table 1.** Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical *Yap1* binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of <i>Yap1</i> site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)	<i>YAP1</i>	Similarity to bacterial <i>csgA</i> protein	10.4
YML007W			Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	<i>ATR1</i>	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	<i>OYE3</i>	NADPH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	<i>OYE2</i>	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to <i>A. thaliana</i> zeta-crystallin homolog	3.7
YOL126C		<i>MDH2</i>	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

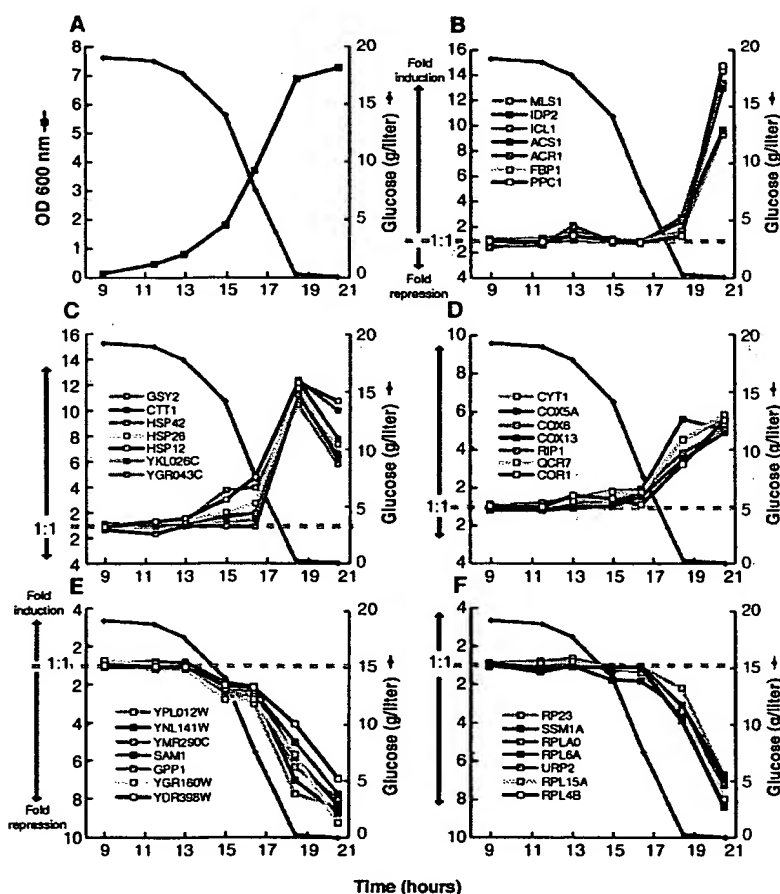
works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

## REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 OFFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100- $\mu$ l PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3 $\times$  standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at [cmgm.stanford.edu/pbrown](http://cmgm.stanford.edu/pbrown). After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratagene). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-



**Fig. 5.** Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contains STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

- tion, the bound DNA was denatured by a 2-min incubation in distilled water at  $-95^{\circ}\text{C}$ . The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.
10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at  $30^{\circ}\text{C}$  with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at  $-80^{\circ}\text{C}$ .
  11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25  $\mu\text{g}$  of polyadenylated [poly(A)\*] RNA, primed by a dT(16) oligomer. This mixture was heated to  $70^{\circ}\text{C}$  for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500  $\mu\text{M}$  for dATP, dCTP, and dGTP and 200  $\mu\text{M}$  for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100  $\mu\text{M}$ . The reaction was then incubated at  $42^{\circ}\text{C}$  for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with 470  $\mu\text{l}$  of 10 mM Tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to  $\sim 5 \mu\text{l}$ , using Centricon-30 microconcentrators (Amicon).
  12. Purified, labeled cDNA was resuspended in 11  $\mu\text{l}$  of  $3.5\times$  SSC containing 10  $\mu\text{g}$  poly(dA) and 0.3  $\mu\text{l}$  of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for  $\sim 8$  to 12 hours in a water bath at  $62^{\circ}\text{C}$ . Before scanning, slides were washed in  $2\times$  SSC, 0.2% SDS for 5 min, and then  $0.05\times$  SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
  13. The complete data set is available on the Internet at [cmgm.stanford.edu/pbrown/explore/index.html](http://cmgm.stanford.edu/pbrown/explore/index.html)
  14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
  15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database ([genome-www.stanford.edu](http://genome-www.stanford.edu)), Yeast Protein Database ([quest7.proteome.com](http://quest7.proteome.com)), and Munich Information Centre for Protein Sequences ([speedy.mips.biochem.mpg.de/mips/yeast/index.html](http://speedy.mips.biochem.mpg.de/mips/yeast/index.html)).
  16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* 14, 3813 (1994).
  17. S. Kratzer and H. J. Schuller, *Gene* 161, 75 (1995).
  18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* 268, 12116 (1993).
  19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* 242, 727 (1994).
  20. A. Hartig et al., *Nucleic Acids Res.* 20, 5677 (1992).
  21. P. M. Martinez et al., *EMBO J.* 15, 2227 (1996).
  22. J. C. Varela, U. M. Praskett, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* 15, 6232 (1995).
  23. H. Ruis and C. Schuller, *Bioessays* 17, 959 (1995).
  24. J. L. Parrou, M. A. Testa, J. Francois, *Microbiology* 143, 1891 (1997).
  25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
  26. S. L. Forsburg and L. Guarente, *Genes Dev.* 3, 1166 (1989).
  27. J. T. Olesen and L. Guarente, *ibid.* 4, 1714 (1990).
  28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* 13, 119 (1994).
  29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B-C, G, or T; N-G, A, T, or C; R-A or G; and Y-C or T.
  30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* 12, 363 (1996).
  31. D. Shore, *Trends Genet.* 10, 408 (1994).
  32. R. J. Planta and H. A. Reue, *ibid.* 4, 64 (1988).
  33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.
  34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* 15, 3187 (1995).
  35. P. Lesage, X. Yang, M. Carlson, *ibid.* 16, 1921 (1996).
  36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin et al., *Mol. Microbiol.* 20, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg et al., *J. Biol. Chem.* 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PFK1* and *PFK2* [P. A. Moore et al., *Mol. Cell. Biol.* 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase *TCT1* [P. H. Bissinger et al., *ibid.* 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas et al., *J. Biol. Chem.* 266, 15602 (1991); U. M. Praskett and P. A. Meacock, *Mol. Gen. Genet.* 223, 97 (1990); D. Wotton et al., *J. Biol. Chem.* 271, 2717 (1996)].
  37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).
  38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero et al., *Yeast* 11, 137 (1995)], *MLS1/DAL7* (73% identical) (20), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
  39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* 11, 3307 (1991).
  40. D. Tzamaras and K. Struhl, *Nature* 369, 758 (1994).
  41. Differences in mRNA levels between the *tup1 $\Delta$*  and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the *tup1 $\Delta$*  strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
  42. The *tup1 $\Delta$*  mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco RI site 124 base pairs before the stop codon of the *TUP1* gene.
  43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* 15, 341 (1995).
  44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* 148, 149 (1994).
  45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* 242, 250 (1994).
  46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* 60, 1783 (1994).
  47. A. Muheim et al., *Eur. J. Biochem.* 195, 369 (1991).
  48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* 269, 32592 (1994).
  49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at [cmgm.stanford.edu/pbrown](http://cmgm.stanford.edu/pbrown). Images were scanned at a resolution of 20  $\mu\text{m}$  per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
  50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).
  51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the *Tup1* deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997